# Systems Analysis of Eukaryotic Proteomic Regulatory Mechanisms

By

Parimal Samir

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry

May, 2016

Nashville, Tennessee

Approved by:

Professor Andrew J. Link

Professor Melanie D. Ohi

Professor Nicholas J. Reiter

Professor Kevin L. Schey

Professor William P. Tansey

To those who believed in me and the giants who lent their shoulders to stand on…

# Acknowledgements

It was an absolute privilege to train and work under the guidance of Dr. Andrew Link. He taught me the value of patience, when I was struggling, and encouraged me to overcome my limitations. He allowed me to hone my skills and gave me independence to pursue ideas that developed my independent thinking. His criticisms, and encouragements, allowed me to grow and hopefully become a better scientist. He gave me opportunities, trust, confidence, and guidance that, I believe, are rarely accorded to a graduate student.

I have had an awesome thesis advisory committee with Dr. Richard Armstrong, Dr. David Hachey, Dr. Melanie Ohi, Dr. Nick Reiter, Dr. Kevin Schey, and Dr. Bill Tansey being members. They have helped me grow through their guidance, criticisms and encouragements. I would start off very nervous in my committee meetings, but I would be calm and relaxed by the end owing to their encouragements. I would not have asked for a better or a more understanding committee.

I would also like to thank all the people who have helped me along the way during the graduate school. I am especially grateful to Dr. David Miller. I was ready to leave science when he helped me get into graduate school. I am also thankful to Dr. Walter Chazin who was willing to let me train under him. I would like to thank Dr. Antonis Rokas and Dr. Daniela Drummond-Barbosa who let me rotate through their labs. I would like to thank Dr. Michelle Grundy and Dr. James Patton for their help and guidance during the IGP years. I would also like to thank my wonderful collaborators; Dr. James Thomas, Dr. Tom Dever, Dr. Charles Spencer, Dr. Sebastian Joyce, Dr.

Mark Boothby, Dr. Barbara Natalizio, Dr. Mei Wei, Dr. Joachim Frank, Dr. Bingxin Shen, Ms. Ming Sun, Dr. Tony Weil, Dr. Cherie Sumanasekera, Dr. Jian Ling, Dr. Leigh Howard, Dr. Kathryn Edwards, Dr. Ralf Krahe, Dr. Bjarne Udd, Dr. Shawn Levy, and Dr. Nripesh Prasad.

I had a wonderful time in the Link lab. One of the reasons was the help and guidance of Link lab members. I would like to especially thank Dr. Chris Browne for all the fun including the, sometimes, heated discussions. All the other lab members helped me grow too. To Ms. Allison Galassie, Dr. Adam Farley, Dr. Morgan Sammons, Dr. Mu Zheng, Dr. Xinnan Niu, Dr. Kristen Hoek, Ms. Tara Allos, and Dr. Laura Gordy – thank you!

I could not have reached here without the sacrifices of my parents. They provided me with opportunity to grow and explore my interests. They kept believing in me, when so few did. They let me pursue my dreams, while I repeatedly failed, shooting for the stars. Always encouraging and gently nudging.

There are only a few people who have had an unwavering belief in my abilities. My kid brother is one of them. For some reason he was always on my case that I was frittering my life away. He was annoying because he would always tell of my escapades to my parents. He would not let any of my foibles go uncommented. I now realize he had higher expectations of me than I had of myself. He was always there to provide support when I needed it the most.

I would like to remember my Grandma, for whom I can do no wrong. I would also like to thank my Grandpa for all the wonderful times. To my cousins Minal, Abhishek,

Rishu, Babbu, Tarunika, Abbu, Ashu, Rohit, and Chhoti, who let me be the big brother to them – you are the best!

The list of people I am indebted to is infinite and I could not do justice to all of them. But I would like to mention one who has been with me through the highs and lows of graduate school – Aditi, my wife. She brought structure and discipline to my life. She has been my biggest critic, and my biggest champion. Her thoughtful advice have helped me grow as a scientist, and a person.  She has stuck with me even though, sometimes, I made it very difficult to do so.

# Summary

I studied three biological problems in my dissertation research. The problems involved flow of information into the cells from outside, the regulation of information flow by the ribosomes in protein synthesis, and the disruption of information flow due to microsatellite repeat expansions leading to a human disease myotonic dystrophy. In the first study, I built a conceptual basis for interpreting and understanding the cellular responses to multiple concurrent stimuli. A gene represents the inherent information of the cells while a stimulus represents the information outside their boundary. Since a gene and a stimulus are both packets of information, they can be considered analogues. Therefore, the concepts of gene interactions can be applied to the study of modulation of cellular processes by stimuli. This assumption allowed me to define the concepts of environmental interactions and environmental epistasis in terms of gene interactions and genetic epistasis. I used proteomic and transcriptomic changes in *Saccharomyces cerevisiae* to test the conceptual framework. In the second study, I designed and performed experiments to test the ribosome filter hypothesis. The ribosome filter hypothesis says that the amount of information flow from a transcript to a protein is regulated by the compositions of the subpopulations of ribosomes in a cell. The composition of a ribosome determines its interactions with the mRNA and accessory factors, which in turn determine the efficiency of translation of a transcript. Therefore, to efficiently translate the proteome required for growing in one environmental condition would require a specific complement of ribosomes with different compositions. The required complement of ribosomes will be different for a cell growing in a different environmental condition. A difference in the protein composition of

ribosomes from cells growing in two different conditions would be evidence supporting the ribosome filter hypothesis. It would allow identification of candidate ribosomal proteins, or their post-translation modifications that regulate information flow from specific transcripts. I used growth of *S. cerevisiae* with fermentable carbon source, glucose, and non-fermentable carbon source, glycerol, as two conditions. I used iTRAQ labeling based quantitative proteomics as well as, in collaboration with the Joachim Frank lab, cryo-electron microscopy to measure the changes in protein composition of ribosomes. I used yeast genetics and polysome profiling to measure the effect of loss of function of a candidate ribosomal protein, Rpl8a or Rpl8b, on translation. In the third project, I studied the changes introduced in the skeletal muscle proteome of myotonic dystrophy patients, both type 1 and 2, due to the disruption of information flow by microsatellite repeat expansions in the non-coding regions of mRNA transcripts. I used iTRAQ labeling based quantitative proteomics analysis to quantitate the changes in the skeletal muscle proteome of DM patients compared to healthy volunteers. I identified differentially present proteins and used pathway analysis to understand their role in the pathogenesis. I have identified a number of candidate proteins that are interesting targets for more in depth genetic and biochemical studies including a ribosomal protein *RPL13A*, previously implicated in regulating information flow by translational inhibition of transcripts containing the GAIT sequence motif. In summary, I have studied three different ways the information content of cells and tissues are affected.

# Table of Contents

# Table of Figures

# Chapter I

## Introduction

An organism alters its biochemical state in response to the changes in its environment or the stage in its life cycle. One way it can alter its biochemical state is by regulating its cellular proteome. Regulation of the cellular proteome is essential for continued survival of an organism. In my research, I studied three facets of regulation of cellular proteomes that I will describe in the subsequent chapters. This introduction has been divided into three parts to reflect the three projects described in my thesis. In the first part, I will present experimental evidence and model the cellular responses to multiple concurrent changes in the environmental conditions. I will describe a conceptual framework that helps understand the biological effects of the concurrent stimuli using changes in the cellular proteome and transcriptome of *Saccharomyces cerevisiae*. In the second project, I will present experimental evidence and discuss the role of ribosome as a regulatory element in translational control of the cellular proteome in *S. cerevisiae*. In the third project, I will dissect the misregulation of the human skeletal muscle proteome due to the expansion of microsatellite repeat elements in the human genome that leads to a human disease myotonic dystrophy.

**Regulation of the proteome by environmental stimuli.**

The interaction of an organism with its environment determines its internal biochemical state. In turn, the organism modifies the biochemical state of its environment by secretion of biomolecules or dissipation of chemical energy. The process by which the organism modifies its biochemical state requires information flow

between the organism and its environment. The information packet could be biomolecules, such as signaling molecules or nutrients, or physiochemical agents such as pH and temperature. The organism uses the information stored in its genetic material as well as its current biochemical state to bring about the required modifications to its biochemical state. The biochemical state is a reservoir of information and a component of the information repertoire of a cell. Other components of the information repertoire include spatial distribution of biomolecules, their chemical structures, and their conformational states.

The proteome is one of the critical components of the biochemical state. In most of the biochemical processes, proteins, the building blocks of a proteome, act as molecular actuators providing essential biochemical activities. The proteome is very dynamic. New molecules are constantly being synthesized and old molecules degraded. Not all proteins are present at the same abundance levels. The cell needs to fine tune its synthesis and degradation machinery to maintain an optimal level of every protein in the molecule. The optimal levels of different proteins depend upon the external environmental conditions, or stimuli.

An important motivation for studying the modification of proteome by environmental stimuli comes from the study of tumors and their microenvironments. The environmental conditions inside the tumor microenvironments are different from the physiological conditions at multiple levels (Vaupel, Kallinowski, and Okunieff 1989; Mbeunkui and Johann Jr 2008; Trédan et al. 2007; Finger and Giaccia 2010; Song 1984; Kessenbrock, Plaks, and Werb 2010; CHUNG et al. 2005; Hazlehurst, Landowski, and Dalton 2003; Kenny, Lee, and Bissell 2007; Whiteside 2008). Tumor

microenvironment has been found to promote tumor growth by activating survival pathways. It also helps tumor cells escape the host immune response (Whiteside 2008). The local tumor microenvironment allows cells in the tumor to crosstalk, which might contribute to continued survival signaling through autocrine loops (Mbeunkui and Johann Jr 2008). It can promote drug resistance by modulating the delivery of a drug or its stability (Trédan et al. 2007). The local tumor microenvironment has also been proposed to provide sanctuaries for subpopulations of tumor cells that facilitates acquisition of drug resistance (Hazlehurst, Landowski, and Dalton 2003). Essentially the tumor microenvironment expands the information repertoire of the tumors allowing the cells in it to escape the host machinery designed to inhibit uncontrolled cell growth.

Another motivation for studying the effect of environmental stimuli comes from the goal of ensuring food security for an ever growing world population (Hanjra and Qureshi 2010; Rosegrant and Cline 2003; Fan et al. 2011; Godfray et al. 2010). To achieve that we need to develop strains of plants that can grow in a wide variety of stress conditions, such as drought, high salinity, and extreme fluctuations in temperatures, to name a few. This would require understanding the cellular responses to environmental stimuli (Chapin III, Autumn, and Pugnaire 1993; Apel and Hirt 2004; W. J. Chen and Zhu 2004; De Angelis and Gobbetti 2004; Mizoguchi, Ichimura, and Shinozaki 1997).

**Microarray based transcriptomic studies:** The advent of high throughput technologies, for example microarrays, heralded a new era in the study of modifications of the information repertoire of the cells in response to environmental stimuli (Schena et al. 1995; Lockhart et al. 1996). Due to technological reasons the initial focus of these

studies was on the transcriptomic responses. One of the earliest systems level studies of the transcriptome involved its modulation by environmental stimuli in human cell culture models. In this study two stimuli were used, heat shock at 43 $^{o}$C for 4 hours and growth in the presence of phorbol esters for 4 hours (Schena et al. 1996). . Comparison of the transcriptomic response to the two stimuli revealed distinct changes specific to them. The technique was later applied to study transcriptomic changes in cancer. A pioneering study identified characteristic changes in the transcriptome that accompanied tumor suppression (J. DeRisi et al. 1996).

Budding yeast *Saccharomyces cerevisiae* was at the forefront of these studies. In one of the earliest studies transcriptomic studies, *S. cerevisiae* was used to study metabolic reprogramming in response to the changes in the nutrient availability (J. L. DeRisi, Iyer, and Brown 1997). A study with large numbers of stimuli revealed that there are common genes and pathways that are activated or repressed in response to all environmental stimuli. The authors called this group of genes the environmental stress response genes (Gasch et al. 2000). Although these studies provided important insights into modifications of the information repertoire of the transcriptome in responses to environmental stimuli, the translation of the insights to the proteome was not straight forward. Development of new technologies was needed for systems level study of the information repertoire of proteomes.

**Mass spectrometry based proteomics:** Mass spectrometry is a very powerful technique for studying chemical composition and structure of molecules. Its utilization was initially limited because of the lack of a technique to ionize and get the biomolecules in gaseous state. Development of matrix assisted laser desorption

ionization (MALDI) and electrospray ionization (ESI) provided a handle to study biomolecules. Application of MALDI was first to be reported and was used to analyze molecules with up to 100 000 m/z (Tanaka et al. 1988). Soon after, the application of ESI to study oligonucleotides and proteins was reported (Fenn et al. 1989). Earlier the application of tandem mass spectrometry in sequencing proteins and peptides had been demonstrated (Hunt et al. 1986). In this study, enzymatic digestion of proteins to yield smaller peptides followed by liquid chromatography fractionation was used. The peptides were ionized using liquid secondary ion mass spectrometry (Hunt et al. 1986). These developments catalyzed rapid explosion in the application of mass spectrometry for studying large biomolecules, including proteins (Aebersold and Mann 2003; Yates, Ruse, and Nakorchevsky 2009).

The mass spectrometers used in proteomics studies have also undergone rapid improvements over the last two decades. It has included improvements in resolution and mass accuracy as well as improved ion optics and data acquisition speeds (Yates, Ruse, and Nakorchevsky 2009; Walther and Mann 2010; Smith 2002; X. Han, Aslanian, and Yates III 2008; Michalski et al. 2011; Senko et al. 2013). These improvements are allowing study of proteomes at ever greater depth (Kim et al. 2014; M. Wilhelm et al. 2014).

A feature of ESI is that ions are generated from a solution. This allows inline coupling of an ESI source to liquid chromatography systems for separation of peptides. Liquid chromatography tandem mass spectrometry is the most widely used technology in proteomics for both protein identification and quantitation (Rudnick et al. 2010).

**Quantitative proteomics approaches:** In the last decade the focus of mass spectrometry based proteomics has shifted towards quantitative studies from the generation of catalogs of protein identifications (Altelaar, Munoz, and Heck 2013; Ong and Mann 2005; Bantscheff et al. 2007; Bantscheff et al. 2012; Larance and Lamond 2015). A number of techniques have been developed for quantitation using mass spectrometry. The techniques can be broadly divided into two categories – (1) label free and (2) labeling based. Alternatively, the techniques can be divided into categories based upon whether the quantitation is done using precursor ions or the fragment ions. In this case too there can be two categories – (1) precursor ion based and (2) fragment ion based.

Label free approaches have been the most popular quantitative approach due to the ease of sample preparation and a reduced cost of running experiments. Examples of label free approach include quantitation using area under the curve of precursor ion intensities, as a peptide is detected during elution from the liquid chromatography column, and spectral counting (Neilson et al. 2011). Label free approaches have lower precision and accuracy that led to the development of a number of isotopic labeling approaches. The isotope labels can be added metabolically as in the stable isotope labeling by amino acids in cell culture (SILAC) and Neutron encoding stable isotope labeling by amino acids in cell culture (NeuCode SILAC) (Hebert et al. 2013; Ong et al. 2002). It can also be added chemically as in isobaric tag for relative and absolute quantitation (iTRAQ), isotope coded affinity tag (ICAT), tandem mass tag (TMT) and mass-coded abundance tagging (MCAT) (Ross et al. 2004; Gygi et al. 1999; Thompson et al. 2003; Cagney and Emili 2002).

TMT and iTRAQ are isobaric tagging approaches in which quantitation is done using the fragment ion intensities (Thompson et al. 2003; Ross et al. 2004). The tags are designed in such a way that the mass added to the tagged peptides is the same across the set of samples being analyzed together. This allows co-isolation and fragmentation of peptides in the mass spectrometers. Peptide fragmentation releases reporter ions whose masses differ from each other. The reporter ion intensity is proportional to the amount of peptide in the samples. Since peptides from all the samples in a set are sampled at the same point in time, comparing the ratios of the reporter ions provides a measure of relative quantitation of the peptides. If a common control sample is used in one of the reporter ion channels, any number of samples can be quantitated relative to a common control (Hoek et al. 2015). Similar experiment designs have been developed with precursor level quantitation approaches for large scale quantitative proteomics approaches(Geiger et al. 2010).

The advances in mass spectrometry based quantitative proteomics have been critical for the systems level studies of the information repertoire of the proteomes. Information in the proteome is encoded through abundances of proteins, their post-translational modifications, and spatial localization of the molecules. Mass spectrometry based proteomics is revolutionizing the research in every aspect of biology (Yates, Ruse, and Nakorchevsky 2009; Stastna and Van Eyk 2012; Clancy and Hovig 2014; Choudhary and Mann 2010; Drissi, Dubois, and Boisvert 2013; Gajadhar and White 2014; Y. Zhang et al. 2013; Hennrich and Gavin 2015; Altelaar, Munoz, and Heck 2013; Breker and Schuldiner 2014).

**Studying protein abundance changes:** The cellular abundance of proteins is one feature of the information repertoire encoded through the proteome. The abundances of specific proteins are changed in response to an environmental stimulus (Feder and Hofmann 1999; Lindquist 1986; Blokhina, Virolainen, and Fagerstedt 2003; Roth, Roepenack-Lahaye, and Clemens 2006). Some of the changes lead to synthesis of proteins that are needed for responding to the stimulus, for example heat-shock proteins upon heat-shock (Feder and Hofmann 1999). Others might lead to a downregulation, for example that of pro-inflammatory receptors to avoid tissue damage (Ohta and Sitkovsky 2001). The goal of systems biology is to understand the contributions of all the components of an organism towards its continued survival and adaptation to new environments (Kitano 2002; Kitano 2000; Ideker, Galitski, and Hood 2001). Understanding the global protein level changes is a minimum requirement towards fulfilling the goals of systems biology. Mass spectrometry based proteomics has been one of the most widely used approaches in this area.

In one of the earliest applications of quantitative proteomics, ICAT was used to study the differences between the steady state proteomes of *S. cerevisiae* cells growing with either ethanol or galactose as carbon source. The differences in expression of two isoforms of alcohol dehydrogenase, *ADH1* and *ADH2* that are 93% identical at the amino acid sequence level, were determined. These differences were similar to the predicted differences based upon their distinct functions in carbon metabolism (Gygi et al. 1999). Soon after another stable isotope labeling approach for quantitation, SILAC, was used to study muscle differentiation in cell culture using C2C12 myoblasts (Ong et al. 2002). Quantitative proteomics is also a popular method to study proteomic changes

in cancer (Ong and Mann 2005; Xu et al. 2008; Everley et al. 2004; Hanash, Pitteri, and Faca 2008; Wulfkuhle, Liotta, and Petricoin 2003). Multiple reaction monitoring approach has been used to determine the abundances of proteins present in 41 copies per cell to more than a million copies per cell (Picotti et al. 2009).

In another study, the proteomic changes with either carbon or nitrogen limitation in *S. cerevisiae* was assayed using N15 labeling in chemostat cultures. This study identified 102 differentially expressed proteins; many of those changes were expected based upon previous studies. The proteins that were upregulated in carbon limitation showed good correlation with the transcriptomic changes. However, the proteins that were upregulated in nitrogen limitation did not correlate well with the transcriptomic changes. This suggested a transcriptional regulation of carbon source limitation response while the predominant mode of regulation in response to nitrogen limitation was either translational or degradation controlled (Kolkman et al. 2006).

A perturbation study, where one of the environmental parameter is perturbed keeping everything else constant, is a powerful technique. It allows us to find the changes in the biochemical state in response to the perturbation. However, cells in their native environment are rarely subject to single discrete changes in their environment. Therefore the application of this approach in modeling complex cellular responses in native conditions is limited. One way to expand the power of this approach would be to perform combinatorial perturbation analysis where multiple environmental stimuli are applied concurrently.

**Cellular responses to combinatorial stimuli**: Combinatorial effects of compounds have been an active area of research in toxicology, drug combination therapy, and environmental science (Greco, Bravo, and Parsons 1995; Altenburger et al. 2013; Altenburger, Nendza, and Schüürmann 2003; Altenburger et al. 2012; Altenburger, Walter, and Grote 2004; Faust et al. 2001; Ankomah and Levin 2012; GARDNER 2002; Berenbaum 1989; Deneer 2000; Schoen 1996; Hermens, Leeuwangh, and Musch 1985). Most of these studies focused on one aspect of the cellular responses such as mixture toxicity or therapeutic effect of a combination of drugs. However, there have been only a limited number of systems level studies of cellular responses to multiple concurrent stimuli. Most of these have been transcriptomic studies.

In a pioneering study of transcriptomic changes in response to combinatorial changes in the environmental stimuli, regression analysis was used to interpret the observed changes. In this study 10 environmental parameters were varied. The total number of unique conditions was 55 and the total number of experiments was 170. Combinatorial stimulus was found to have profound effect on the expression pattern (Knijnenburg et al. 2009). The same analysis approach was also used in a study of transcriptomic responses of *Arabidopsis thaliana* liquid cultures to concurrent stimuli. In this study, the stimuli were high salinity and carbon dioxide concentration (Kanani, Dutta, and Klapa 2010). Although the regression models were able to explain significant amount of the statistical variances in the two studies, a biological interpretation was not straightforward.

In a more recent study, the combinatorial effects of high NaCl and pheromone signaling was assayed in *S. cerevisiae* (Vaga et al. 2014). Phosphorylation events were

used as measures of activation/repression of specific signaling pathways. A set of ordinary differential equations were used to build logic models describing the integration of signaling through the high osmolarity and mating signaling pathways. This approach identified complex interconnections between the two pathways. However, similar to the regression approach above, a biological interpretation of the data is not straightforward. This motivated me to explore approaches that can be used to precisely and accurately model the combinatorial responses and at the same time have a simpler biological interpretation.

I searched upon the conceptual framework of gene interactions that can be easily applied to the study of environmental stimuli (P. C. Phillips 1998). As an abstraction a gene is a packet of information, so is an environmental stimulus. A gene has an effect on the information repertoire of the cell when the information stored in it is used to build functional molecules such as regulatory RNAs or proteins. In cells the products of multiple genes carry the information from their respective genetic loci. The information from genes travels through the information networks inside cells. The interaction between the gene products, or sometimes a lack of interaction due to a loss of function, integrate the information and modify the information repertoire of the cells. The information for modification comes from within the cells, its genetic material. In the case of an environmental stimulus, the information from outside the cell travels through the cellular information network to modify its information repertoire. If multiple stimuli are present, the information from each of them would be integrated inside the cellular information network.

As an abstraction, this integration of information could be similar to the integration of information from genes through gene interactions. Therefore, the conceptual framework of gene interactions can be used to study the combinatorial effect of concurrent environmental stimuli. I called it the concepts of environmental interactions and environmental epistasis. I defined an environmental interaction as the interaction between different environmental stimuli that affect the same observable characteristic or trait. In this schema, environmental epistasis is a special case of environmental interaction in which the effects of the individual stimuli are not independent of each other. We have tested the applicability of this approach in studying the effects of multiple concurrent environmental stimuli in *S. cerevisiae* (Samir et al. 2015).

**Regulation of the proteome by the ribosomes**

The information stored in the genome is the template used to build the information repertoire stored in a proteome. It involves encoding of the information into an intermediate class of molecules called messenger RNAs (mRNAs) followed by translation of the information from mRNAs into amino acid sequences of proteins through a process called translation. Translation consists of 4 steps: (1) initiation, (2) elongation, (3) termination, and (4) recycling. A host of proteins and RNAs regulate these steps (Kapp and Lorsch 2004). The catalytic engine of this process is the ribosome. The eukaryotic ribosome consists of a 60S large ribosomal subunit and a 40S small ribosomal subunit. The large subunit has three rRNA molecules (28S, 5.8S and 5S) and 46 ribosomal proteins. The small subunit is made up of a single rRNA (18S) and 33 ribosomal proteins (Jonathan R Warner 1999; Nakao, Yoshihama, and

Kenmochi 2004). The catalytically competent fully functional 80S ribosome is a heterodimeric complex of the small and large subunits.

Ribosomes are assembled inside the nucleus through a process called ribosome biogenesis. Ribosome biogenesis is one of the most energy intensive processes. In eukaryotes requires concerted action of hundreds factors, including proteins and small nucleolar RNAs (snoRNAs) (J. Woolford 2015; J. L. Woolford and Baserga 2013; de la Cruz, Karbstein, and Woolford Jr. 2015; Turowski and Tollervey 2014; Planta 1997; Boisvert et al. 2007; J. R. Warner 1989). RNA Pol I transcribes the pre-rRNA whose endolytic processing generates 28S, 18S, and 5.8S rRNAs. RNA Pol III transcribes 5S rRNA (Kressler, Linder, and Cruz 1999; Venema and Tollervey 1999; Granneman and Baserga 2004; Nazar 2004). During ribosome biogenesis, pre-RNA is processed and the ribosomal proteins sequentially added (Gamalinda et al. 2014). Once the ribosomal subunits have been assembled, they are exported out of the nucleus and undergo a final round of processing before joining the free ribosomal subunit pool primed to start translation (Johnson, Lund, and Dahlberg 2002; Rouquette, Choesmel, and Gleizes 2005; Zemp and Kutay 2007; van Riggelen, Yetil, and Felsher 2010).

In the first step of translation, translation initiation factors help assemble a functional ribosome on an mRNA (Kapp and Lorsch 2004). Eukaryotic mRNAs contain 5’ cap structure in which a guanosine nucleotide is connected through 5’-5’ bond. This guanosine is also methylated on position 7. The cap acts as the start beacon, among its many functions (Shatkin 1976). It helps recruit eIF4 initiation factors to the mRNA. The eIF4 complex unwinds the secondary structures on the mRNA and helps recruit the 43S preinitiation complex (PIC) (Gingras, Raught, and Sonenberg 1999, 4). PIC consists of

40S ribosomal subunit bound to eIF2 and initiator Met-tRNA. eIF2 in PIC is bound to GTP. The 40S subunit scans the mRNA to find the initiation AUG codon. The 40S ribosome starts scanning the mRNA to find the initiation AUG codon. Once the initiation codon has been identified, the GTP bound to eIF2 is hydrolyzed and eIF2 dissociates from the complex. This paves way for recruitment of the 60S ribosomal subunit leading to the formation of a fully functional ribosome and translation elongation can begin (Hinnebusch 2005; Kapp and Lorsch 2004; Sonenberg and Hinnebusch 2009; Gingras, Raught, and Sonenberg 1999; Korostelev 2014).

The cap-dependent mode of translation initiation discussed above is the predominant mode of translation initiation. In addition, a cap-independent translation initiation mechanism can also be employed by some mRNAs (Merrick 2004). In cap independent translation initiation, the ribosomes are recruited directly to an internal ribosome entry site (Merrick 2004; Pelletier and Sonenberg 1988; Jang et al. 1988; Chappell, Edelman, and Mauro 2000). Some viruses exploit this mechanism of translation by shutting down the cap-dependent translation that shuts down most of the host protein synthesis. This allows viral protein synthesis to occur using a cap-independent mechanism (Sk et al. 1989; Firth and Brierley 2012; Boehringer et al. 2005; Fernández et al. 2014).

Once the 80S ribosome is assembled on an mRNA, translation elongation can begin (Kapp and Lorsch 2004). This phase of translation requires two elongation factors, eEF1 and eEF2, and tRNAs charged with cognate amino acids. tRNAs are the keys for decoding the information from mRNA. They contain three letter anticodon key that is complementary to the three letter codons on mRNA. The elongation factor eEF1

facilitates the entry of charged tRNA to the free ribosome acceptor site as well GTP hydrolysis upon correct anticodon-codon base pairing. Upon correct anticodon-codon base pairing, a new peptide bond is formed to extend the length of the nascent polypeptide through peptidyl transferase reaction. The peptidyl transferase catalytic activity resides in the 28S rRNA component of the 60S large ribosomal subunit, making ribosome an example of ribozyme. Another molecule of GTP is consumed by eEF2 elongation factor for translocating the ribosome three nucleotides once the peptidyl transfer reaction has occurred. This ensures that information from the mRNA is decoded sequentially three nucleotides at a time. GTP hydrolysis helps in reducing the errors during translation elongation as well as provides directionality (Kapp and Lorsch 2004; G. R. Andersen, Nissen, and Nyborg 2003; Nilsson and Nissen 2005; Nyborg and Liljas 1998; Frank 2012).

Translation elongation continues till the ribosome encounters one of the stop codons - UAA, UAG, and UGA. The stop codons mark the end of the message in the mRNAs. There are no tRNAs for any of the stop codons. Instead, the translation termination factor eRF1 is recruited to the free acceptor site on the ribosome followed by the binding of another termination factor eRF3 (Kapp and Lorsch 2004; Dever and Green 2012). The termination factor eRF1 can recognize all of the three stop codons. Once it has ensured that the ribosome has reached a stop codon, it catalyzes the peptide release from the ribosomes. Although peptide release can be catalyzed by eRF1 alone, presence of eRF3 greatly increases the reaction rate. The cooperative actions of eRF1 and eRF3 ensure proper and speedy translation termination. After peptide has been released the complex of tRNA, termination factors, ribosome, and the

mRNA is known as post-termination complex (Dever and Green 2012; Georges et al. 2014; Inge-Vechtomov, Zhouravleva, and Philippe 2003, -).

The final step in translation is recycling of the ribosomes from the post-termination complex to start another round of translation. This is also the least understood phase in translation, especially in eukaryotes. There are two fates possible for the ribosome bound to the mRNA after termination – (1) it can reinitiate translation at a downstream start codon, or (2) it can be dissociated from the mRNA making it available for another round of translation. Both possibilities have been found to be utilized in cells, albeit the latter being more frequent than the former (Nürenberg and Tampé 2013; D. J. Young et al. 2015; S. K. Young et al. 2015; Dever and Green 2012; Jackson, Hellen, and Pestova 2012; Franckenberg, Becker, and Beckmann 2012).

In the traditional model of translational control, all powers to regulate are invested in the protein and RNA accessory factors, and the ribosomes are considered passive players. Most studies of translational control have focused on such regulatory molecules. This effort has led to the identification of the core components of translational control and a better understanding of this process (Kapp and Lorsch 2004; Dever and Green 2012; Sonenberg and Hinnebusch 2009; Hinnebusch 2015). In recent years, the idea that the ribosomes are regulatory elements in gene expression regulation has been gaining ground. The role of ribosome in the regulation of gene expression is a very active area of research (Ruggero and Pandolfi 2003; Kondrashov et al. 2011; Xue et al. 2015; Jonathan R. Warner and McIntosh 2009; Jonathan R. Warner 2015; McIntosh and Warner 2007; Komili et al. 2007). This idea was nucleated

due to the observation of heterogeneity in the ribosomal sub-populations in cellular slime molds (S. Ramagopal 1992).

**Ribosome heterogeneity**: Heterogeneity in ribosome composition has been known since the early work with *E. coli* ribosomes (Kurland et al. 1969). Soon after studies in rat revealed the differences between the protein composition of ribosomes from muscle and liver (Sherton and Wool 1974). Studies in cellular slime mold *Dictyostelium discoideum* revealed three modes of heterogeneity (S. Ramagopal 1992). In the first mode, some ribosomal proteins were found to be exclusively present in the vegetative or differentiated state. Two of the ribosomal proteins were present only in the vegetative state, while three only in the differentiated spores (S. Ramagopal 1992; Subbanaidu Ramagopal and Ennis 1981). In the second mode, some ribosomal proteins were present in varying stoichiometries in different developmental stages (Subbanaidu Ramagopal and Ennis 1981; S. Ramagopal 1992). These modes of generating ribosomal heterogeneity were found to be conserved across several cellular slime mold species (Subbanaidu Ramagopal and Ennis 1984). A third mode of heterogeneity generation involved differential post-translational modifications of ribosomal proteins. In *Dictyostelium discoideum,* ribosomal proteins were found to be methylated or phosphorylated. Many of the modifications were specific to vegetative cells or starvation induced aggregation competent cells (S. Ramagopal 1992; S. Ramagopal 1991). The fact the heterogeneity was present in growth stage specific manner suggested a functional significance of heterogeneity.

In a more recent study in *A. thaliana*, liquid chromatography tandem mass spectrometry analysis of protein composition of ribosomes revealed changes in

composition in response to sucrose feeding. Some of the changes in composition involved paralogous ribosomal proteins. Paralogous ribosomal proteins may have arisen through gene or whole genome duplications and are nearly identical to each other. In *A. thaliana,* 231 ribosomal protein genes code for ribosomal proteins. Each ribosome contains 79 proteins (Hummel et al. 2012). The presence of paralogs provides a mechanism of generating ribosomal heterogeneity through a specific use of paralogs in *A. thaliana* ribosomes (Nakao, Yoshihama, and Kenmochi 2004).

In *S. cerevisiae* too there are 59 ribosomal protein paralog pairs (Nakao, Yoshihama, and Kenmochi 2004). Therefore, the presence of paralog can allow heterogeneity in the ribosomal populations also in *S. cerevisiae*. In a study involving oxidative stress, *RPL22A* and *RPL16B* were found to be upregulated in response to hydrogen peroxide treatment (Chan et al. 2012). A mutational analysis revealed that loss of *rpl22a* but not *rpl22b* leads to sensitivity to the oxidative stress. The other paralog pair, *RPL16A* and *RPL16B,* did not show sensitivity to oxidative stress suggesting a functional complementation between them (Chan et al. 2012). The lack of functional complementation between *RPL22A* and *RPL22B* suggest a functional role for paralog mediated ribosomal heterogeneity in translational control.

**The ribosome filter hypothesis**: The ribosome filter hypothesis was proposed in light of the observations that some mRNAs contained regions that were similar or complementary to 18S or 28S rRNAs (Vincent P Mauro and Gerald M Edelman 2007). It was expanded to include ribosomal protein mediated interactions with the mRNAs. The underlying idea is that the ribosomes with different compositions translate specific mRNAs with differing efficiency (Figure 1) (Mauro and Edelman 2002; Vincent P Mauro

and Gerald M Edelman 2007). There are four basic tenets of the hypothesis that I am reproducing here – *"(1) the ribosome is a regulatory structure that embodies mechanisms for preferentially translating different subsets of the message population, (2) ribosomes may display a continuum of regulatory effects, (3) competition for binding sites in ribosomal subunits may affect the rate of translation of different mRNAs, and (4) the filter may also be modulated as a result of altering or masking particular binding sites on ribosomes"* (Mauro and Edelman 2002; Vincent P Mauro and Gerald M Edelman 2007).

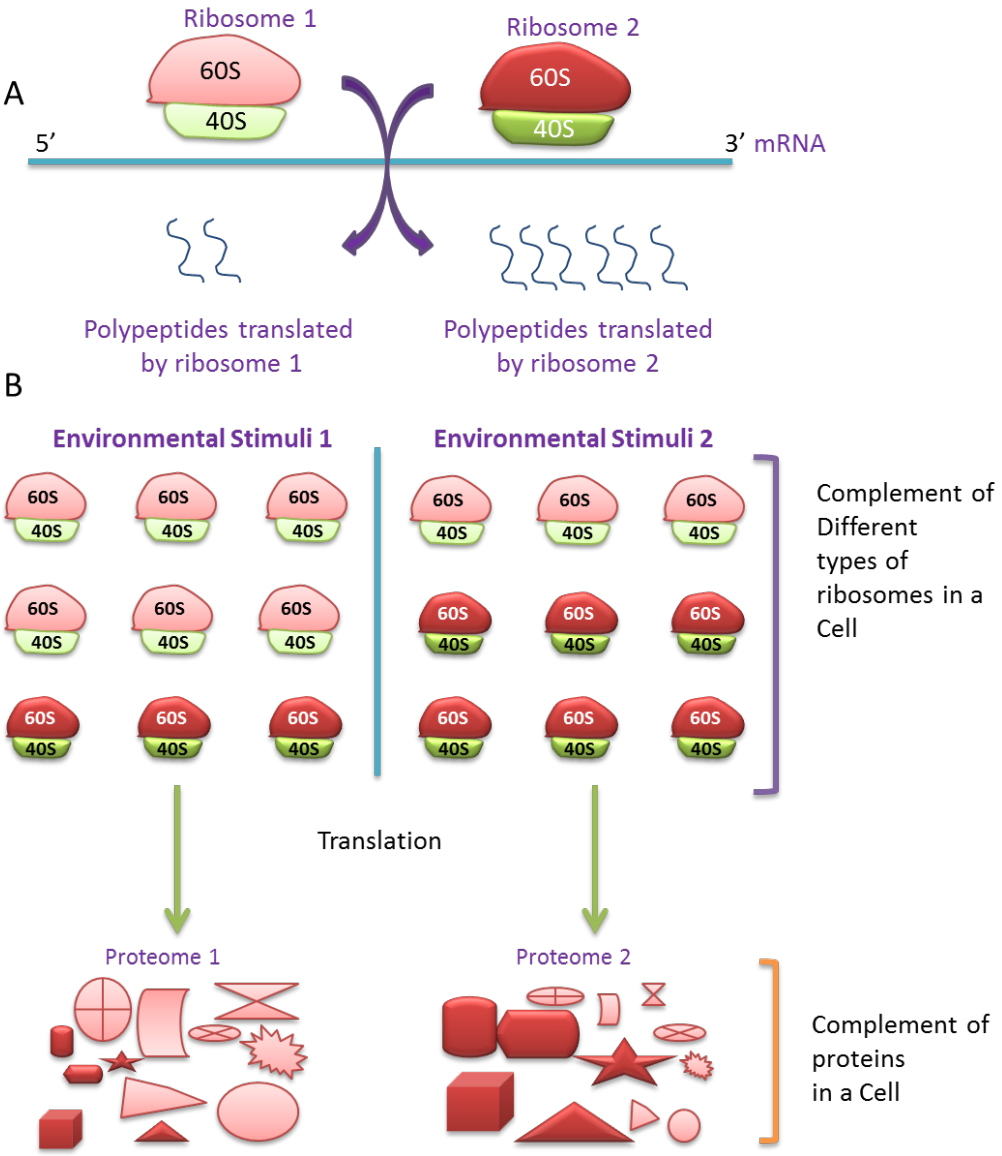



**Figure 1 -** The ribosome filter hypothesis.

A) A cartoon showing that two ribosomes in differing composition translate a transcript with different efficiencies. The dark ribosome is more efficient than the light ribosome. B) A cartoon showing the presence of heterogeneous subpopulations of ribosomes differing in their composition. To efficiently translate two different proteomes, two different subpopulations of ribosomes are needed.

**The depot hypothesis**: The depot hypothesis was proposed to explain the extra-ribosomal functions of ribosomal proteins (Ray, Arif, and Fox 2007). In one of the earliest examples of extra-ribosomal function of a ribosomal protein, RPL13A was found to inhibit translation of ceruloplasmin (Cp) mRNA (Mazumder et al. 2003). RPL13A is a component of interferon-Gamma Activated Inhibitor of Translation (GAIT) complex that binds to GAIT element in the 3'UTR of Cp mRNA (Mazumder et al. 2003; R. Mukhopadhyay et al. 2009). Under normal conditions, RPL13A remains associated with the 60S subunit of the ribosomes. It is phosphorylated in response to interferon-Gamma singling. Phosphorylation is the trigger for its dissociation from the ribosome. The free phosphorylated protein binds to the GAIT element to inhibit translation of target mRNAs (Mazumder et al. 2003). This observation led to the authors *"to propose a 'depot hypothesis' in which macromolecular assemblies, while maintaining their ordinary activity, acquire the non-canonical capability to release component proteins that perform new functions outside the complex"* (Ray, Arif, and Fox 2007).

A number of examples of ribosomal proteins have been described in addition to the one described above (Jonathan R. Warner and McIntosh 2009). In mammals, S27-like and S27 have been found to inhibit MDM2 mediated ubiquitination of p53 (Xiong et al. 2011). In *A. thaliana*, phosphorylation of RPL10 causes it to translocate to nucleus where it is hypothesized to be involved in host defense against virus infection (Carvalho et al. 2008). These examples suggest a frequent utilization of ribosomal proteins in functions outside their primary function in translation as predicted by the depot hypothesis. The ribosome filter and depot hypothesis have a human disease element too.

**The ribosomopathies**: Ribosomopathies have been defined as "*a collection of disorders in which genetic abnormalities cause impaired ribosome biogenesis and function, resulting in specific clinical phenotypes*" (Narla and Ebert 2010). The earliest example of a ribosomopathy is a human disease Diamond Blackfan Anemia (DBA) (Jonathan R. Warner 2015; Narla and Ebert 2010). Mutations in *RPS19* locus were found to be associated with DBA. The types of mutation included nonsense, frameshift, splice site, and missense mutations (Draptchinskaia et al. 1999). Since then mutations in 11 ribosomal protein genes have been found to be associated with DBA (Aspesi et al. 2014).

Since discovery of mutations in ribosomal proteins in DBA, a number of other ribosomopathies have identified. One example is X-linked Intellectual Disability caused by the mutation in RPL10. The mutant protein is still functional and is able to complement the conditional mutant of RPL10. The missense mutation also led to an increase in actively translating ribosomes (Zanni et al. 2015). Mutation in RPS20 was found to cause predisposition to Hereditary Nonpolyposis Colorectal Carcinoma (Nieminen et al. 2014). The list of ribosomopathies is growing with new mutations and diseases continuously being found to be associated with each other (Narla and Ebert 2010; Jonathan R. Warner 2015; Yelick and Trainor 2015; Danilova and Gazda 2015; Amanatiadou et al. 2015; Brooks et al. 2014; Ruggero and Pandolfi 2003; Yang et al. 2015; Martin et al. 2014). These studies are providing added motivation for studying the regulatory functions of ribosomes.

In this study, I used changes in protein composition of *S. cerevisiae* ribosomes in response to a change in carbon source to identify candidate ribosomal proteins that

might be playing a role in ribosome filter mediated translational control. I used iTRAQ based quantitative proteomics and cryo-EM, in collaboration with the Joachim Frank lab, to study the changes in composition of ribosomes both at the population as well as single particle levels. I identified 11 such proteins that includes a paralog pair, Rpl8a and Rpl8b, using quantitative proteomics. I am using yeast genetics, biochemistry, and next-generation sequencing in follow up experiments to dissect the exact mechanism of ribosome filter mediated translational control.

**The proteome in myotonic dystrophy**

Myotonic dystrophy (dystrophia myotonica, DM) is an autosomal dominant progressive multisystemic disorder caused by expansion of microsatellite repeats (Thornton 2014). It is the most common form of adult muscular dystrophy (Udd and Krahe 2012). DM was first described by Steinert, Batten and Gibb in 1909 (Schoser and Timchenko 2010). It was later recognized as a multisystemic disorder with the observation of a high incidence of cataracts in DM patients (Schoser and Timchenko 2010). DM is characterized by progressive muscle weakness, myotonia, cataracts, and cardiac conduction defects (Udd and Krahe 2012; Thornton 2014; Machuca-Tzili, Brook, and Hilton-Jones 2005; Turner and Hilton-Jones 2014).

The causative agent for DM was identified as an expansion of CTG repeat element in the 3’ untranslated region of dystrophia myotonica-protein kinase (Brook et al. 1992; Mahadevan et al. 1992; Y. H. Fu et al. 1992). Shortly afterwards, many cases of DM were described that lacked the CTG expansion (Thornton, Griggs, and Moxley 1994; Ricker K et al. 1995; Ricker et al. 1994; Udd et al. 1997; Meola et al. 1996). These patients typically had a milder form of the disease. Since the proximal muscles

are involved in this form of the disease, instead of the distal muscles in the previously described form, it was named proximal myotonic myopathy. The two forms of the disease were later renamed as myotonic dystrophy type 1 (DM1) and myotonic dystrophy type 2 (DM2). DM1 patients had the CTG repeat expansions. DM2 was subsequently found to be caused by a CCTG tetranucleotide repeat expansion in the first intron of *Znf9* gene.

**Myotonic dystrophy type 1 (DM1)**: DM1 is the classic form of the disease that was first described more than a hundred years ago. DM1 can present in four different forms: adult onset, congenital, childhood onset, and late onset oligosympotomatic. The age of onset negatively correlates with the repeat size. The adult form is the most typical form while the congenital form is the most severe form of DM1 (Udd and Krahe 2012). DM1 is more prevalent in European populations, where it ranges from 1:1100 in Finland to 1:10700 in other countries (Thornton 2014). It was found to be a rare disease in Taiwan and sub-Saharan Africa. In Taiwan the disease incidence was found to be 0.46:100000 (Hsiao et al. 2003). Incidences of DM1 in Africa in non-European populations were very rare. For example, only one DM1 family was identified in Nigeria, when the population of country was 120 million. The same study revealed that DM1 was more prevalent in Europe, Japan, Southwest Asia and India. It was less prevalent or extremely rare in West African, Bantu, Ethiopian, Tunisian Berbers, Southern Chinese, Thai and non-European Australians (Ashizawa and Epstein 1991).

The molecular basis of DM1 was independently identified by three groups (Mahadevan et al. 1992; Y. H. Fu et al. 1992; Brook et al. 1992). Interestingly, one of the studies mistakenly identified the repeat to be a GCT repeat. They were using GCT

repeat synthetic oligonucleotides as probes in their experiment. An oligonucleotide with repeated GCT sequences is virtually identical to a CTG repeat sequence, except the ends, that confounded their interpretation (Y. H. Fu et al. 1992). Previous studies had identified the long arm of chromosome 19 to contain the locus that was in linkage disequilibrium with DM1 (Korneluk et al. 1989; Brunner et al. 1989; Shaw et al. 1986; Smeets et al. 1990; Aslanidis et al. 1992). Other two studies used a positional cloning strategy to clone the previously identified DM1 locus and identified the CTG repeat expansions as the causative agent (Brook et al. 1992; Mahadevan et al. 1992). The number of repeats in normal individuals was found to vary between 5 and 30. In DM1 patients, the number of repeats exceeded 50 (Brook et al. 1992; Mahadevan et al. 1992).

CTG repeats are highly unstable. In one study only 4 out of 110 cases were identified where the repeats were passed on unchanged from the parental generation. In the same study, the repeat expansions by more than 400 in a single generation were observed (Redman JB et al. 1993). The repeats are also biased towards expansion (Thornton 2014; Redman JB et al. 1993; Temmerman et al. 2004). There is also a pronounced maternal expansion bias (Pearson, Edamura, and Cleary 2005). Different mechanisms for expansions have been proposed that include genome duplication errors, genome maintenance error in quiescent state or recombination defects during meiosis (Pearson, Edamura, and Cleary 2005).

A high amount of somatic instability in CTG repeat length suggests a prominent role for errors in genome maintenance or duplications that occur after meiosis (Loreto Martorell et al. 1997). A comparison between the repeat lengths of identical twins

showed different patterns of expansion in one of the two pairs (de Munain et al. 1994). The repeat was also found to expand with time in the same individual (L. Martorell et al. 1995). A number of studies have reported variable length of the repeats in different tissues of the same individuals (Thornton, Johnson, and Moxley 1994; Shelbourne et al. 1992; Lavedan et al. 1993; G. Jansen et al. 1994). Interestingly, repeat contraction was reported in a father-son pair with repeat length falling to the normal range for the son who did not have the disease (Shelbourne et al. 1992). In one of the studies, CTG repeats were found to be larger in skeletal muscles compared to leukocytes (Thornton, Johnson, and Moxley 1994). Taken together, it suggests a more prominent role for errors in genome maintenance during mitosis or quiescence in repeat expansion.

**Myotonic dystrophy type 2 (DM2)**: DM2 was described shortly after identification of CTG repeat expansion as the cause for classical DM. Although DM1 and DM2 are similar in their symptoms, and being a RNA dominant disorder, there are important differences between the two. DM2 is a relatively milder form of DM. DM2 does not involve a severe central nervous system defect. It is primarily a late onset disease. A congenital form of DM2 has not been observed. (Thornton 2014; Machuca-Tzili, Brook, and Hilton-Jones 2005; Turner and Hilton-Jones 2014, -; Cho and Tapscott 2007; Ulane, Teed, and Sampson 2014; Ranum and Day 2002). The numbers of tetranucleotide CCTG repeats are below 30 for normal individuals while they vary between 75 and 11000 in affected individuals making it much larger than the ones observed in DM1 (Liquori et al. 2001). Surprisingly, the repeat length has not been found to be associated with the age of onset or severity of the disease (Ranum and Day 2002).

In normal individuals CCTG repeats are interrupted by GCTG motif, TCTG motif, or both (Kurosaki et al. 2012; Liquori et al. 2003). The repeat tracks are polymorphic with the form (TG)n(TCTG)n(CCTG)n(NCTG)n(CCTG)n (Bachinski et al. 2009). In affected individuals, only the CCTG motif of the repeat expands (Kurosaki et al. 2012; Day et al. 2003; Liquori et al. 2001; Bachinski et al. 2009). The repeats are very instable. They can contract as well as expand over generations (Kurosaki et al. 2012; Day et al. 2003; Ulane, Teed, and Sampson 2014). Nuclear magnetic resonance spectroscopy showed that the CCTG repeats are prone to forming metastable hairpin and dumbbell structures. The structures were shown to undergo dynamic conformational exchange. Both the structures were also found to contain flexible stem (Lam et al. 2011). In another study, it was found that the DM2 repeats are recombination hotspots. This process might be driven by DNA repair mechanisms (Dere and Wells 2006). The repeat itself was proposed to have originated from an insertion of Alu elements in the *ZNF9* gene. The normal repeat structure was found to be conserved in primates, mouse, and rat (Kurosaki et al. 2012; Liquori et al. 2003).

In contrast to DM1, which has a relatively broader geographical distribution, DM2 seems to be more prevalent in European Caucasians. DM2 is more common in Northern European ancestry. Extensive haplotype analysis suggested that DM2 spread from a common founder (Liquori et al. 2003). More recently, a DM2 patient was identified in Japan. Haplotype analysis suggested that the DM2 repeat in this individual originated separately from those in European populations (Saito et al. 2007). There have been incidences of DM2 in non-European individuals in Morocco, Algeria, Lebanon, Afghanistan and Sri Lanka (Saito et al. 2007).

**Molecular mechanisms behind DM pathogenesis**: DM is classified as a RNA-dominant disease (Osborne and Thornton 2006). The disease is thought to be caused by sequestration of essential RNA binding proteins by toxic RNA that contains the repeat elements in their non-coding region. Although this is now a universally accepted mechanism with certain modification, it was not always so clear. In an aptly titled review, *"Myotonic dystrophy: will the real gene please step forward! "*, Sarah Harris, Colin Moncrieff, and Keith Johnson expressed the frustrations of researchers in finding the molecular mechanism of the pathology (Harris, Moncrieff, and Johnson 1996). This was years after identification of CTG repeat expansion as the disease causing agent. A number of putative mechanisms had been proposed that explained some aspect of pathogenesis but not all. In addition, a new type of DM, DM2, had recently been found that did not contain the CTG repeat expansions.

A number of pathogenic mechanism models have been proposed, many of them might contribute to the disease (J. E. Lee and Cooper 2009). Shortly after identification of the repeats, a haploinsufficiency model was proposed. In this model, the decrease in the amount of *DMPK* protein leads to pathogenesis. This model was supported by the observation that *DMPK* protein and mRNA levels were decreased in DM1 patients (Y.-H. Fu et al. 1993). Follow up studies in mouse models suggested that although the decrease in *DMPK* protein levels might be contributing to the disease; it was not the sole cause. In one of the study, the mice null for *Dmpk* had only mild phenotype (Gert Jansen et al. 1996). In another study, *Dmpk* null mice developed late onset myopathy, but did not show all the abnormalities observed in DM1 patients (Reddy et al. 1996).

Furthermore, a loss of function mutation in *DMPK* has not yet been identified suggesting that a decrease in *DMPK* is not the sole cause of DM1 (J. E. Lee and Cooper 2009).

*DMPK* locus lies in a very gene rich region of the genome. Using this information it was proposed that the repeat expansion is affecting the expression of adjacent genes. In one of the studies condensed chromatin was found downstream of *DMPK* gene (Otten and Tapscott 1995). This suggested that the transcription of genes in the vicinity would be decreased. This was indeed found to be a case where the expression of a candidate homeodomain gene *DMAHP* (also known as SIX5) was reduced in DM1 patients. However in a study in mice, knocking out *Six5* led to development of cataract without apparent abnormalities in the skeletal muscles (Klesert et al. 2000). This suggested that the effect of repeat expansion on the chromatin and transcription of neighboring genes might be a contributing factor, not the primary one in DM1 pathogenesis.

The failure of other models to explain all the clinical features of DM1 lead to the proposition of RNA dominance model (Osborne and Thornton 2006). It is also called RNA gain-of-function and RNA toxicity models (Sicot and Gomes-Pereira 2013; J. E. Lee and Cooper 2009). Mice models with CTG repeats in the genome displayed the repeat instability including expansions and contractions in both germline as well as somatic tissues (Monckton et al. 1997; Gourdon et al. 1997). Transgenic mice with DM1 region of a patient inserted in its genome displayed many of the associated pathological features. The pathological features included myotonia, progressive weakness of skeletal muscles, testicular atrophy as well as cognitive dysfunction among others (Seznec et al. 2001). In a cell culture model, expression of *DMPK* cDNA containing 46 CTG repeats

inhibited myoblast differentiation (Usuki et al. 1997). These studies strongly suggested a critical role for RNA in DM1 pathogenesis.

Stronger evidences were provided by experiments in which the CTG repeats were expressed independent of *DMPK*. In a cell culture model, expression of CTG repeat alone was able to inhibit myoblast differentiation (Bhagwati, Shafiq, and Xu 1999). In another cell culture experiment, expression of *DMPK* 3'UTR with expanded repeats was able to delay myoblast differentiation and the normal *DMPK* 3'UTR did not have an effect (Amack, Paguio, and Mahadevan 1999). A transgenic mice model with 250 CTG repeats construct inserted in the first intron of human skeletal actin (HSA-CTG) was able to reproduce most of the disease features (Mankodi et al. 2000). Further evidence was provided by the identification of CCTG repeat as the disease causing agent in DM2.

DM1 and DM2 have similar phenotypes, but are caused by mutations at very different loci in the genome suggesting an involvement of a common pathway. *DMPK* and *ZNF9* have not been identified to function in the same molecular pathway. The mutations in both of the diseases are in the non-coding region of the transcript. Even if the repeats are assumed to be translated, the resulting proteins will have expansions containing different amino acids. Combined with the observation that expression of mutant RNA repeat alone in cell culture models as well as mouse models is able to replicate many of the clinical features of the diseases, a critical role of RNA gain of function seems to be beyond doubt.

**Effects of RNA gain-of-function on splicing**: The expanded CTG repeat caused a number of abnormalities including sequestering of RNA binding proteins in the nucleus, formation foci in the nucleus, changing the methylation state of the surrounding areas, formation of heterochromatin in the adjoining areas, decreasing transcription of the adjoining genes, and a decrease in the amount of *DMPK* protein itself. The diverse array of the effects of CTG expansion confounded the quest for discovery of molecular pathogenesis for a long time till unequivocal evidence in favor of RNA gain-of-function were obtained.

RNA gain-of-function model needs trans acting factors whose misregulation leads to the splicing defects observed in DM. CUG-BP1 was found to bind to CTG repeats in *in vitro* experiments with cytoplasmic and nuclear extracts (L. T. Timchenko et al. 1996). CUG-BP1 is a member of CELF family of proteins with six members (Osborne and Thornton 2006; Ranum and Cooper 2006). The EDEN-BP, a *Xenopus* homolog CELF homolog, was found to be involved in deadenylation (Paillard 1998). In another experiment, CUG-BP1 was found to regulate alternative translation initiation in *in vitro* experiments with mammalian cell extracts. The use of alternative initiation sites is one of the mechanisms to generate different isoforms of C/EBPβ transcription factor (N. A. Timchenko et al. 1999). CUG-BP1, another CELF family member, was found to modulate C to U RNA editing in mammalian cell extracts (Anant et al. 2001). In another study, CELF family members were found to regulate alternative splicing in cell and developmental stage specific manner (Ladd, Charlet-B, and Cooper 2001). In one study, CUG-BP2 was found to bind ARE elements in 3’ UTR of cyclooxygenase-2 mRNA. The binding of CUG-BP2 stabilized the mRNA, but also inhibited translation (D.

Mukhopadhyay et al. 2003). These studies suggested that sequestration of CELF family proteins by the repeat containing toxic RNA could be one mechanism for DM pathogenesis.

However, two very important observations contradicted the model with CELF family proteins at the center of DM pathogenesis mechanism. First, binding of CUG-BP was not found to be proportional to the CUG repeat length (Michalowski et al. 1999). Second, CUG-BP failed to colocalize with the nuclear RNA foci (Fardaei et al. 2001). This suggested that although CELF family proteins may be important in DM, they were not sequestered in the RNA foci and were not the primary splicing regulators important for pathogenesis. In a recent study, CUG-BP1 was found to be overexpressed in skeletal muscle biopsies of DM1 patients but not in DM2 patients (Cardani et al. 2013). This further ruled out CELF family proteins as the common splicing regulator affected in the two types of DM.

The hunt for splicing regulator sequestered by CUG repeat containing RNAs led to the muscleblind-like (MBNL) family of proteins. There are three MBNL proteins in humans (Ranum and Cooper 2006). *Muscleblind* was identified in *Drosophila* as developmentally regulated gene that was important for eye and muscle development (Begemann et al. 1997; Artero et al. 1998). MBNL was identified as a candidate after it was found to bind to the larger CUG expansions *in vitro*. In crosslinking experiments, MBNL was not found to bind to RNA that had less than 11 CUG repeats (Miller 2000). This provided a model in which MBNL proteins were not binding to the normal repeats, but were only binding to the disease causing expanded repeat.

The expression pattern of *MBNL* was found to extensively overlap with the expression *DMPK* (Kanadia et al. 2003). All of the three MBNL proteins were also found to colocalize with the nuclear RNA foci in both DM1 and DM2 cells (Fardaei et al. 2002; Miller 2000; Mankodi et al. 2003). MBNL proteins were found to regulate alternative splicing of cardiac troponin T (cTNT) and insulin receptor (IR). Interestingly, CELF family proteins and MBNL promoted different splicing events in cTNT and IR suggesting that they have distinct roles in DM pathogenesis (Ho et al. 2004).

MBNL sequestration can have wide ranging effects on the biology of the cells. In embryonic stem cells, MBNL has been found to negatively regulate expression of pluripotency genes. Knockdown of *MBNL* led to expression of pluripotency genes that were under control of FOXP1 transcription factor (H. Han et al. 2013). In *Mbnl* knockdown mice, fetal tau isoform expression and Mapt isoform misregulation was detected suggesting defects leading to mis-expression of developmentally regulated genes in adults. In DM1 patients there is downregulation of a chloride channel CLCN1. This downregulation is caused by introduction of premature stop codons in the open reading frame (ORF) due to splicing defects (Mankodi et al. 2002; Charlet-B. et al. 2002; Osborne and Thornton 2006). CLCN1 had been shown to be the primary cause of myotonia (Koch et al. 1992). Interestingly, overexpression of CUG-BP was able to recapitulate the aberrant splicing defect in CLCN1 (Charlet-B. et al. 2002). Since CELF family proteins and MBNL proteins have opposite effects on determining the splicing pattern, could it be that the disruption of equilibrium between the two is the primary driver behind splicing defects in DM (Charlet-B. et al. 2002; Ho et al. 2004)?

MBNL1 has also been found to be involved in biogenesis of miR-1 in heart muscles from DM patients (Rau et al. 2011). In rats, miR-1 defect has been shown to be involved in heart development and its misregulation leads to heart conduction defect (Zhao et al. 2007). Regulation of miR-1 biogenesis by MBNL1 explains the heart conduction defects observed in DM patients.

Transcriptomic studies in mouse models have revealed that the majority of defects in DM1 can be explained by the loss of MBNL1 function. In the study, mRNA expression of three mice strains were compared; (1) a transgenic mice expressing CUG repeat, (2) *Mbnl* knockout mice, and (3) a *Clcn1* null mice (Osborne et al. 2009). In another study, both CUG-BP1 and MBNL1 were found to bind to the 3'UTR of target mRNAs and promote mRNA decay (Masuda et al. 2012). Comparison of mouse models expressing expanded CUG repeat containing mRNA or defective *Mbnl1*, revealed that more than 80% of the splicing defects can be explained by the loss of MBNL1 function (Du et al. 2010). A study in mouse and *Drosophila* models has revealed a global role for MBNL proteins in regulating the localization of mRNA (Wang et al. 2012). Taken together, MBNL proteins lie at the center of pathogenesis mechanism in DM.

**Differences in mechanism between DM1 and DM2**: Pathogenesis mechanism in both DM1 and DM2 were thought to be solely mediated by RNA gain-of-function. A contradiction arose with studies in mouse models of DM2. In contrast to *Dmpk,* loss of function of *ZNF9* led to multisystemic defects in *ZNF9* heterozygous mice that was reminiscent of DM2 (W. Chen et al. 2007). This suggested critical role for *ZNF9* protein in DM2 pathogenesis. In yeast *S. cerevisiae*, *ZNF9* was found to be a constituent of ITAF complex that regulates cap independent translation (Gerbasi and Link 2007).

Although the initial studies did not reveal a change in *ZNF9* protein and mRNA levels in DM2 patients, many subsequent studies found a decrease in both *ZNF9* protein and mRNA levels (Udd and Krahe 2012). In myoblasts derived form DM2 pateints, IRES mediated translation was found to be decreased. *ZNF9* was also found to directly bind the IRES elements in the 5' UTR of orinithine decarboxylase mRNA and activate cap-independent translation (Sammons et al. 2010).

These studies have revealed very important insights in the pathogenesis mechanisms of DM1 and DM2. Transcriptomic studies have shed light on the changes in mRNA abundances as well as RNA processing. However, the effect on the global proteome is poorly understood. This study is expected to shed light on the changes in the proteome of DM patients as well as mouse models.

# Chapter II

## Environmental interactions and epistasis are revealed in the proteomic responses to complex stimuli


### Abstract

Ultimately, the genotype of a cell and its interaction with the environment determine the cell's biochemical state. While the cell's response to a single stimulus has been studied extensively, a conceptual framework to model the effect of multiple environmental stimuli applied concurrently is not as well developed. In this study, we developed the concepts of environmental interactions and epistasis to explain the responses of the S. cerevisiae proteome to simultaneous environmental stimuli. We hypothesize that, as an abstraction, environmental stimuli can be treated as analogous to genetic elements. This would allow modeling of the effects of multiple stimuli using the concepts and tools developed for studying gene interactions. Mirroring gene interactions, our results show that environmental interactions play a critical role in determining the state of the proteome. We show that individual and complex environmental stimuli behave similarly to genetic elements in regulating the cellular responses to stimuli, including the phenomena of dominance and suppression. Interestingly, we observed that the effect of a stimulus on a protein is dominant over other stimuli if the response to the stimulus involves the protein. Using publicly available transcriptomic data, we find that environmental interactions and epistasis regulate transcriptomic responses as well.

## Introduction

In their native environments, cells continuously respond to a complexity of environmental stimuli. These include ambient temperature fluctuations, nutrient availability, signaling molecules, and physical forces. In response, cells adjust their biochemical state through multiple mechanisms including the differential production, modification, and degradation of transcripts and proteins (Gasch et al. 2000; Gerner et al. 2002; Pratt et al. 2002; Soufi et al. 2009; Yan et al. 2006). Both extracellular signaling and the metabolic environment strongly influence a cell's growth and responses to therapeutic treatments (Whiteside 2008; Vaupel, Kallinowski, and Okunieff 1989; Trédan et al. 2007; Hazlehurst, Landowski, and Dalton 2003). Model organisms have been used extensively to study cellular responses to individual and combinations of environmental stimuli (Gasch et al. 2000; Brauer et al. 2008; Nicola et al. 2007; Kanani, Dutta, and Klapa 2010; Knijnenburg et al. 2009; Knijnenburg et al. 2007; Murray et al. 2004; Tai et al. 2005; Vaga et al. 2014). We extend these approaches by developing and testing a novel conceptual framework to study proteomic responses of cells to the combinatorial effects of multiple concurrent environmental factors. We have modeled our analysis of these complex environmental interactions using the concepts of gene interaction and genetic epistasis.

Gene interaction is defined as the interaction between genes at different loci that affect the same characteristic or a trait (Pierce 2005). Classically, genetic epistasis is referred to a type of gene interaction in which a mutation at one locus masks or suppresses the phenotype of a mutation at a different locus (Pierce 2005; Bateson 1909). To test the independence of the effects of individual genes, genetic epistasis has

also been defined mathematically as a type of gene interaction in which the combined effect of two or more mutations is not the sum of the effects of the individual mutations (Cordell 2002; P. C. Phillips 2008; Fisher 1958).

Conceptually, the problem of studying multiple concurrent environmental stimuli is similar to the problem of studying the effects of multiple genetic mutations. The product of a gene functions as part of one or more functional modules in concert with the products of many genes. The changes in a gene, for example its loss of function or gain of function, affects the phenotype due to the changes in the activity of the functional modules. If multiple genetic alterations are present, the total effect is due to the integration of the effects of the individual alterations through the functional modules. Similarly, environmental stimuli affect the biochemical state of the cells through specific sensing, signaling, and response modules. Concurrent application of multiple environmental stimuli, similar to the genetic alterations, requires the integration of information from these modules to mount an optimal response. By considering an environmental stimulus as an analogue of a gene, we hypothesized that the concepts of gene interaction and epistasis can be extrapolated to devise a conceptual framework for studying the combined effects of multiple concurrent stimuli. There are several benefits of using this approach; (1) all the genetic, biochemical, and computational tools and concepts developed for studying gene interactions would become available for studying the effects of the environment, (2) it would allow for easier mechanistic interpretation of the responses to complex environmental stimuli, (3) the contributions of an individual stimulus to altering biological processes can be more easily elucidated, and (4) it would

provide a unifying framework for studying gene-gene, gene-environment and environment-environment interactions.

In this study, we define an environmental interaction as the interaction between different environmental stimuli that affect the same observable characteristic or trait. Similar to the statistical definition of genetic epistasis, environmental epistasis is an environmental interaction in which the effects of the individual stimuli are not independent of each other (Cordell 2002; Fisher 1958; P. C. Phillips 2008). To test our hypothesis, we used the yeast S. cerevisiae and grew cells at standard conditions (glucose, 30$^{o}$C) and changed growth conditions to either high temperature (37$^{o}$C, HT stimulus) or the non-fermentable carbon source glycerol (G stimulus) and concurrently with both environmental stimuli (glycerol, 37$^{o}$C, HT+G stimuli) (Figure 2). Using precise quantitative proteomics of the S. cerevisiae proteome and the changes in protein abundance as the readouts of the interactions, we show that environmental interactions and epistasis play central roles in determining the state of the proteome in response to multiple, concurrent environmental stimuli. We also show that, using the dominance of one stimulus over another, environmental interactions can be used to identify proteins that are important for responding to a dominant stimulus. We validated our approach using an independent publicly available transcriptomic dataset.

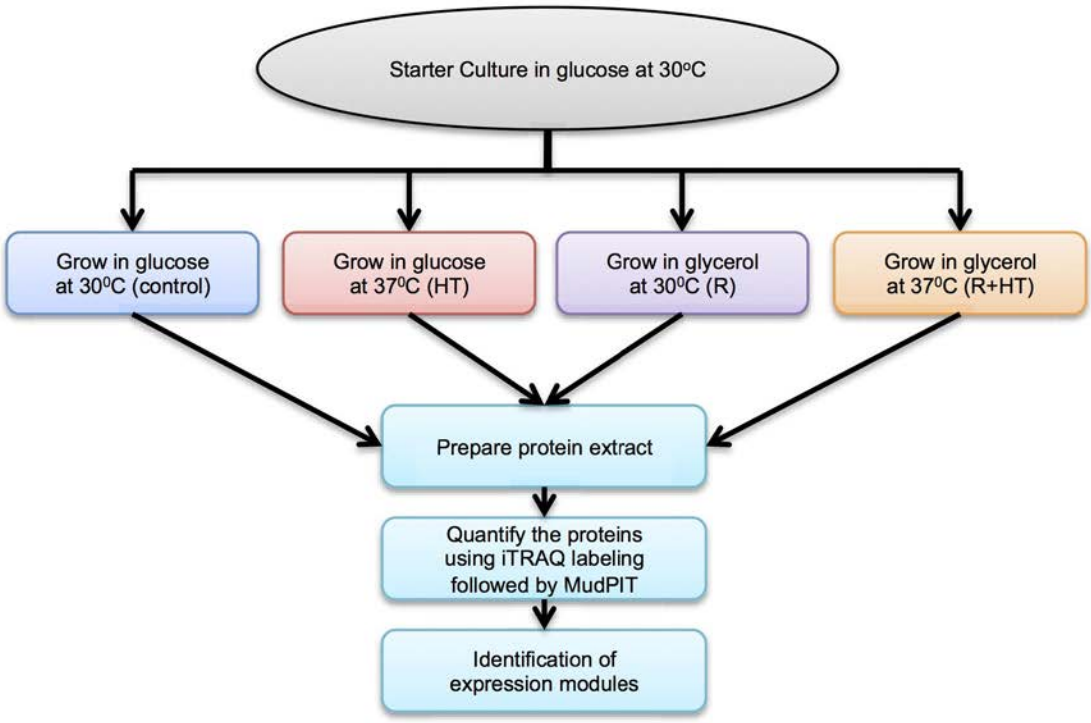



**Figure 2 -** Experiment design to study environmental interactions and epistasis.

Experimental design workflow used in this study. Two environmental stimuli used were high temperature and glycerol as the carbon source. Diploid S. cerevisiae cells (BY4743) were grown in rich media under 4 conditions: 1) glucose at 30$^{o}$C (used as control), 2) glycerol at 30$^{o}$C (G stimulus), 3) glucose at 37$^{o}$C (HT stimulus), and 4) glycerol at 37$^{o}$C (HT+G stimuli). Three biological replicates were performed for each condition.

## Materials and Methods

**Strains and Media.** All experiments used the diploid S. cerevisiae strain BY4743, which has been previously described (Baker Brachmann et al. 1998). Cells were grown using standard techniques (Amberg, Burke, and Strathern 2005).

**Growth rate analysis**. Cells were grown in 96 well plates in 100 µL cultures (10 µL of starter culture and 90 µL of fresh media) with continuous shaking in a BioTek Synergy™ 4 Hybrid Microplate Reader for 10 h. Growth rates were assayed at 8 conditions: (1) Synthetic complete medium with glucose (ScD) at 30$^{o}$C, (2) ScD at 37$^{o}$C, (3) Synthetic complete medium with glycerol (ScG) at 30$^{o}$C, (4) ScG at 37$^{o}$C, (5) Yeast extract, peptone medium with glucose (YPD) at 30$^{o}$C, (6) YPD at 37$^{o}$C, (7) Yeast extract, peptone medium with glycerol (YPG) at 30$^{o}$C, and (8) YPG at 37$^{o}$C. Absorbance was measured at 660nm at 3 min intervals. Using custom R scripts, the doubling times were calculated from the linear regression curve through the log growth phase using the log of the absorbance and time of growth. A two-tailed t-test of independence with Bonferroni correction for the 11 comparisons (7 comparisons of the control, YPD at 30$^{o}$C, to the test conditions, 3 comparisons of the observed concurrent double stimuli effect to the expected sum of individual stimulus effects, and 1 comparison of the observed concurrent three stimuli effect to the expected sum of the effects of the three individual stimulus) was used to calculate the statistical significance of a stimulus effect on the growth rate (Dunn 1961).

**Preparation of yeast protein extracts**: Five mL of YPD (1% yeast extract, 2% peptone, 2 % glucose) was inoculated with a single yeast colony from a YPD agar plate and grown overnight. Three replicates were grown under each growth condition: YPD

at 30°C and 37°C and YPG at 30°C and 37°C. 50 mL of YPD was inoculated with 50 µL of the overnight culture and grown at 30°C and 37°C. 100 mL of YPG (1% yeast extract, 2% peptone, and 3% glycerol v/v) was inoculated with overnight cultures and grown at 30°C and 37°C. The cultures were grown with constant shaking at 175 rpm in Innova 44 shaker incubators (New Brunswick Scientific). For all four growth conditions, cells were harvested at mid-log phase as determined by OD600 measurements. Cells grown in YPD were harvested after 14 h, while cells grown in YPG were harvested after 24 h. All cultures were centrifuged at 2000 rpm for 5 min at 4°C using a Sorvall HLR6/H600A/HBB6 rotor in Sorvall RC-3B centrifuge and washed with ice cold deionized H20. The cell pellets were resuspended in 1 mL ice cold wash buffer (10 mM Tris pH 8.0, 5 mM beta-mercaptoethanol, 500 mM ammonium chloride, 100 mM magnesium acetate) and lysed at 4°C using glass beads and a Bead Beater (BioSpec, Inc) for 10 min as previously described (Browne et al. 2013). The whole cell extracts (WCE) were clarified by centrifugation at 20,000g for 15 min at 4°C, and a 200 µL aliquot of the cleared WCE was stored at -80 °C.

**iTRAQ labeling**: The total protein concentration was determined using a Bradford assay according to the manufacturer's protocol (Sigma Aldrich). For each growth condition, 50 µg of total protein was mixed with 50 ng of bovine serum albumin (Thermo Scientific) as an internal standard. Each protein sample was acetone precipitated and resolublized in 25 µl iTRAQ dissolution buffer (500 mM triethylammonium bicarbonate, 0.1% sodium dodecyl sulfate). The proteins were reduced with tris(2-carboxyethyl)phosphine at 60°C for 60 min and the cysteines were derivatized with methyl methanethiosulfonate at room temperature for 10 min. All samples were

digested with sequencing-grade modified trypsin (1:50; Promega Corporation) overnight at 37°C. Equal fractions of the tryptic digests from the three replicates grown in YPD at 30°C were pooled separately and used as a control for the iTRAQ experiments. Fifty µg of the pooled control and 50 µg of each of the replicates were used for iTRAQ labeling. The iTRAQ labeling reagents were resolublized in 150 µL anhydrous ethanol (Sigma Aldrich). 75 µL of iTRAQ reagent solutions were added to each 50 µg sample, incubated with shaking for 1 h at room temperature on an Eppendorf Thermomixer R, pooled, frozen, lyophilized, resolublized in 200 µL of buffer A (0.1 % formic acid), and stored at -80°C.

**Liquid chromatography and mass spectrometry**: The iTRAQ-labeled samples were analyzed with MudPIT as previously described (Hoek et al. 2015). Briefly, 11 fractions corresponding to ammonium acetate pulses of 25mM, 50mM, 75mM, 100mM, 150mM, 200mM, 250mM, 300mM, 500mM, 750mM, and 1M concentrations were analyzed on 2 hour reverse phase gradients. Precursor ions were analyzed in the Orbitrap mass analyzer followed by four CID fragment ion scans in the ion trap and four HCD fragment ion scans (normalized collision energy = 45%) in the Orbitrap. Dynamic exclusion was enabled with exclusion window of 180 seconds. Monoisotopic precursor selection was enabled.

**iTRAQ data analysis**: RAW files generated by the MudPIT experiments were searched using the Sequest HT database search engine running under Proteome Discoverer v1.4 (Thermo Scientific) against a forward and reverse yeast protein database (S.cererevisiae_orf_trans_all_SGD.fasta.6718) with appended common contaminant sequences (Eng et al. 2008; Eng, McCormack, and Yates 1994). The CID and HCD

spectra were merged using the Spectrum Grouper function in Proteome Discoverer by setting the retention time window to 0.05 minutes and precursor mass tolerance to 10ppm. Beta-methylthiolation of cysteines, and iTRAQ modification of lysine and N-terminus were included as constant modifications. Oxidation of methionine and tryptophan, and deamidation of glutamine and asparagine were used as variable modifications. Precursor mass tolerance was set to 3 Da and fragment mass tolerance was set to 0.8 Da. Protein assembly, reporter ion quantitation, and protein fold change calculations were done using ProteoIQ at 5% peptide and protein FDR (Premier Biosoft). Hierarchical clustering analysis was done using Cluster 3.0 (Eisen et al. 1998). Heatmaps were generated using Java Treeview (Saldanha 2004). Circos plots were generated as described in Krzywinski *et. al.* to visualize the genomic locations of the quantitated proteins (Krzywinski et al. 2009). For better visualization, only those regions of the genome that were quantitated in this study are shown.

**Environmental interaction analysis**: All analyses were performed using R scripts to parse the fold change expression data to identify proteins that show specific expression patterns in response to complex environmental stimuli. For each protein, we used linear regression to test for any association of high temperature or glycerol using a model that included main effects for glycerol and temperature and the glycerol by temperature interaction. We used the effect size estimates and ANOVA p-values (3 df) calculated by the lm function and adjusted the p-values for a 5% FDR using the Benjamini-Hochberg procedure for finding differentially expressed proteins (Benjamini and Hochberg 1995). We used the adjusted *p*-value cut-off of 0.05 to determine statistical significance. If the overall adjusted p-value was greater than 0.05, we classified the proteins as non-

responders. The positive and negative signs of the effect size estimates correspond to upregulation and downregulation, respectively, showing the direction of change. The remaining proteins were further classified into environmental interaction classes based upon the effect size estimate *p*-values and the direction of change. If the *p*-value of an estimate was less than 0.05, the protein was considered differentially expressed in response to that environmental stimulus.

To test if a protein is affected by environmental epistasis, the effect size estimates for the individual high temperature (HT) and glycerol stimuli (G) were summed, the combined standard error calculated as the square root of the sum of the squared standard errors, and a two-sample *t*-test of independence was used to compare the summed effect size estimate to the effect size estimate for the concurrent high temperature and glycerol stimuli (HT+G). If a *t*-test *p*-value was less than 0.05, the protein was assumed to be affected by environmental epistasis.

**Environmental interaction analysis of transcriptomic dataset**: Normalized expression data described in Knijnenburg et. al. was downloaded (Knijnenburg et al. 2009). The transcriptomic data were generated using haploid S. cerevisiae (CEN.PK113-7D MATa) cells grown in chemostat cultures (Knijnenburg et al. 2009). We chose 4 culture conditions similar to our experimental design for further analysis. The culture conditions tested were: 1) with ammonium sulfate as the nitrogen source (n=5), 2) with methionine as the nitrogen source (n=3), 3) anaerobic conditions (n=4), and 4) with methionine as the nitrogen source and anaerobic conditions concurrently (n=3). Transcriptomic data from the cells grown with ammonium sulfate as the nitrogen source were used as the baseline control. The fold change was calculated by subtracting the

average normalized expression data of baseline samples from the individual expression data. Finally, the genes were classified into various types of environmental interaction as described above.

**Co-expression network analysis**: Sparse PArtial Correlation Estimation (SPACE) was used to build protein co-expression networks and identify the hub genes (Peng et al. 2009). To account for outliers, the data were normalized using probabilistic quotient normalization and scaled using a generalized logarithmic scaling factor (Dieterle et al. 2006; Durbin et al. 2002). The data were scaled and centered to have a standard deviation of 1 and mean of 0 to remove any bias in the correlation analysis (Berg et al. 2006). We estimated the partial correlation matrix using the space.dew method implemented in the SPACE R package (Peng et al. 2009). We selected the default value of the tuning parameter for constructing the initial network (Peng et al. 2009). The network was visualized in Cytoscape 3.1.1 (Shannon et al. 2003).

## Results

While cells measure and respond to many environmental stimuli, we chose temperature and carbon source to test our hypothesis. Both stimuli are known to be important factors for survival and have wide-ranging effects on yeast metabolism (Gasch et al. 2000). We used growth with glucose at 30°C as the control, and high temperature and glycerol as the stimuli. The changing growth conditions were: glucose at 37°C (HT stimulus), glycerol at 30°C (G stimulus), and glycerol at 37°C concurrently (HT+G stimulus). To precisely measure the proteomic responses of the cell, we used isobaric tag for relative and absolute quantitation (iTRAQ) labeling followed by multi-dimensional protein identification technology (MudPIT)-based mass spectrometry to

quantify the steady state proteomes under the four different growth conditions (Link et al. 1999; Ross et al. 2004). A total of 1064 proteins were quantitated in the control and the three test conditions. We filtered the data to focus only on the 466 proteins that were quantitated in all three independent replicates of all of the three test conditions (Fig. 3A).

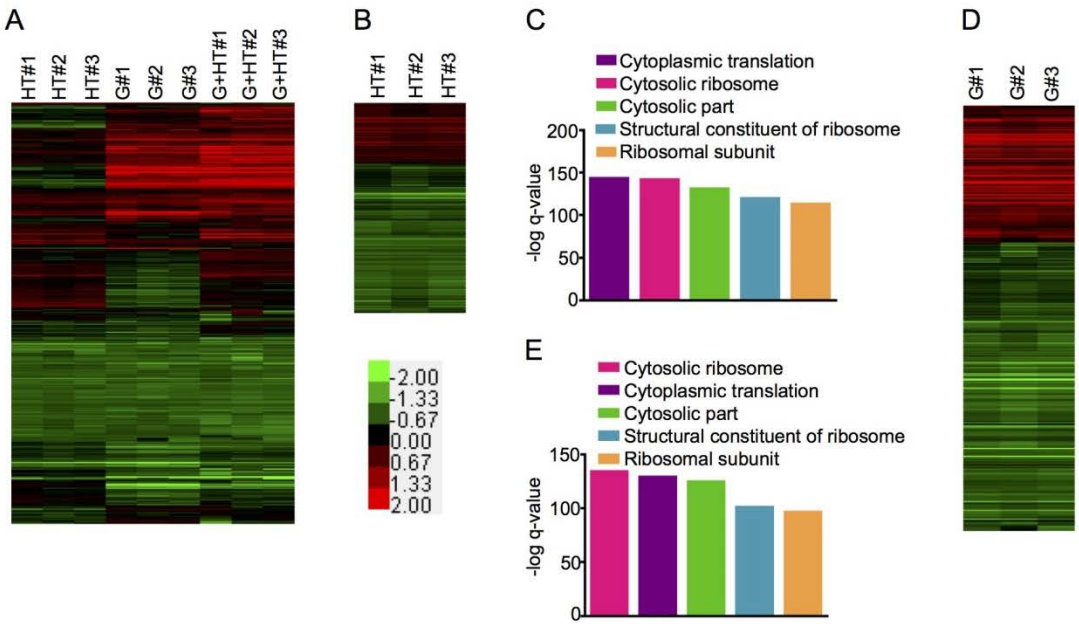



**Figure 3** - Proteomic responses to complex environmental stimuli.

Diploid S. cerevisiae (BY4743) cells were grown in rich media under 4 conditions: 1) glucose as the carbon source at 30$^{o}$C, 2) glycerol as the carbon source at 30$^{o}$C, 3) glucose at 37$^{o}$C, and 4) glycerol at 37$^{o}$C. Three biological replicates for each growth conditions were performed. Fold changes were calculated from iTRAQ reporter ion intensities using reporter ion intensities from the pooled replicates of growth in glucose as the carbon source at 30$^{o}$C as the baseline. The fold changes were $\log_2$ transformed for downstream analysis. The color bar shows the fold change ranges. A) Complete filtered proteomic dataset for high temperature stimulus (HT), glycerol stimulus (G), and concurrent glycerol and high temperature stimuli (HT+G) (Red: Up, Green: Down, Black: No change). The heatmap represents the fold changes of 466 proteins. B) Fold changes of 283 proteins differentially expressed in response to HT stimulus. C) Bar graph shows the –log q-value of enrichments of the top 5 pathways in the list of proteins

differentially expressed after the HT stimulus. D) Fold changes of 379 proteins differentially expressed in response to the G stimulus. E) Bar graph shows the –log q-value of enrichments of the top 5 pathways in the list of proteins differentially expressed after the G stimulus.

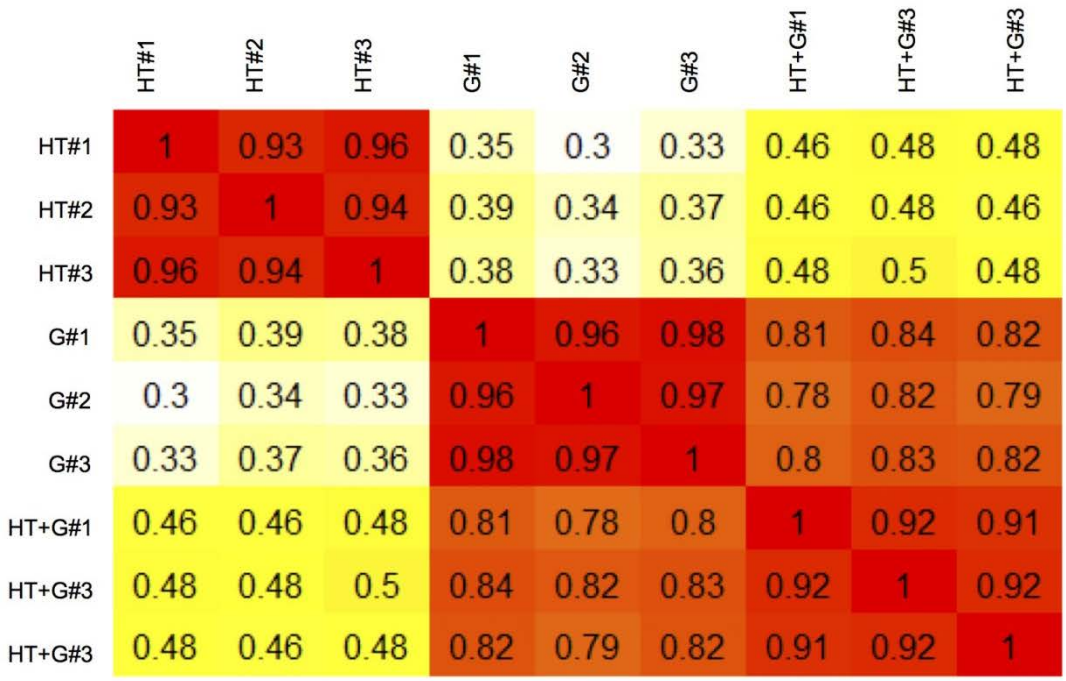



**Figure 4** - Correlation matrix heatmap

Correlation matrix was generated in R. There is a high correlation among replicates showing reproducibility across experimental replicates (red is high).

Cross-correlation analysis of the filtered data showed high reproducibility among the replicates (Figure 4). The proteomic changes in the cells grown with the concurrent stimuli were more similar to the changes induced by glycerol compared to high temperature (Figure 4).

We defined the response to an environmental factor(s) as the log2-fold change in protein abundance/expression between the control and experimental conditions. For this study, we use “fold change” to denote the log2 fold change. We built linear regression models for each protein using fold changes to estimate the effect sizes of the stimuli. We used ANOVA for estimating statistical significances since we were

comparing multiple stimuli. We interpreted the positive or negative sign of the effect size as either upregulation or downregulation, respectively. The Benjamini-Hochberg procedure was used to adjust the ANOVA p-values at 5% FDR (Benjamini and Hochberg 1995). A protein was assumed to be differentially expressed if the adjusted overall ANOVA p-value was less than 0.05. These proteins were further analyzed and classified into different environmental interaction classes using the direction of the change (upregulated or downregulated) and the p-values of the effect size estimates (Benjamini and Hochberg 1995).

**Stimuli-specific expression patterns can be used to identify proteins important for responding to the stimuli.**

We observed 283 proteins differentially expressed with high temperature, 379 proteins differentially expressed in response to glycerol, and 370 proteins were differentially expressed in concurrent high temperature and glycerol (Fig. 3B and D, and Table 1), while 41 proteins did not change in response to any of the stimuli. We selected GeneMANIA Cytoscape plugin for pathway analysis since it extends the input list of differentially expressed proteins by adding related proteins to enhance sensitivity and coverage (Montojo et al. 2010; Mostafavi et al. 2008). It also allows using the complete proteome as the background. This helped to build a more complete picture of differentially regulated pathways. Pathway analysis of these two differentially expressed protein groups revealed the same top five pathways; none were specific to either stimulus (Fig. 3C and E). All of the top five pathways were related to protein synthesis and translational control, suggesting that the regulation of protein synthesis is an important step in responding to environmental stimuli. Translation factors are some of

the most abundant proteins in yeast and our proteomic assays are limited by the abundance of proteins in the cell. Although this could have confounded pathway analysis and led to the identification of translation associated pathways as being the most enriched, using only the differentially expressed proteins suggests that these pathways are, at the least, being differentially regulated. Furthermore, similar observations have also been made in previous studies (Gasch et al. 2000; Brauer et al. 2008; Roberts and Hudson 2006). It is noteworthy that the pathways expected to be important for responding to these stimuli, such as “protein folding” for growth at high temperature and “TCA cycle” for growth with glycerol were present farther down the list at numbers 39 and 53, respectively (Tables 2 and 3) (Richter, Haslbeck, and Buchner 2010; Riezman 2004; Schüller 2003). This mirrors a common problem in ‘omics’ studies that generate large lists of candidate genes, transcripts and proteins. The important responders are lost in a long list where a majority of differentially expressed genes or proteins is not directly responding to the stimulus. Therefore, choosing candidates for an in-depth mechanistic study becomes a challenge.

To address this problem, we devised a methodology using dominance in environmental interactions to identify proteins and pathways important for responding to a stimulus. We noticed proteomic expression patterns in which the response to one stimulus was dominant over the other. We speculated that a protein critical in responding to a stimulus will respond to that stimulus even when challenged by a competing stimulus. If this hypothesis is correct, such an environmental interaction could serve as a filter to select and identify proteins that respond specifically to the dominant environmental stimulus.

To test this hypothesis, we classified the list of 466 proteins responding to the concurrent glycerol and high temperature stimuli based upon their expression patterns. Two classes of dominant environmental interactions are possible. In one class, a stimulus reverses an expression change induced by the other stimulus (Fig5A and B, top panels, rows 1 and 3). In the other class, a stimulus induces a change in expression, while the other stimulus has no effect on its own and does not change the response to the concurrent stimulus (Fig. 5A and B top panels, rows 2 and 4). Each class is represented by two theoretical expression patterns for a total of four expression patterns for each stimulus (Fig. 5A and B top panels).

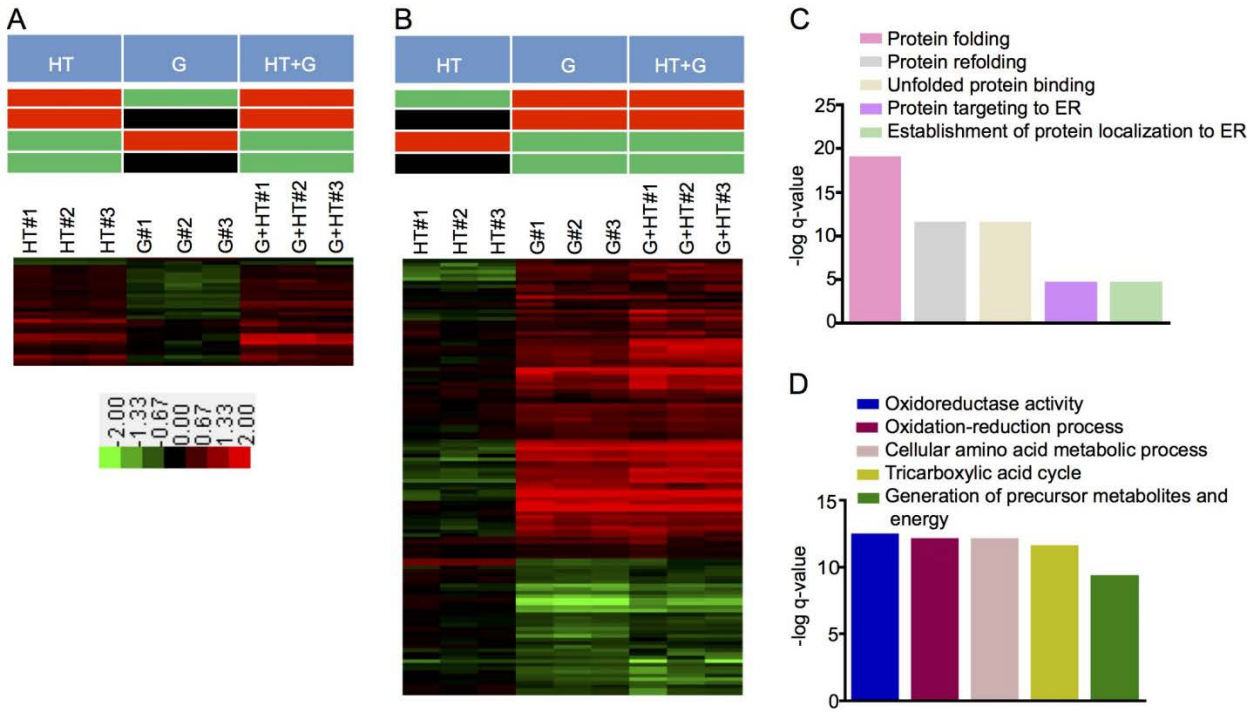



**Figure 5 -** Dominance of an environmental stimulus used to identify proteins that are important for responding to the environmental stimulus

The color bar shows the range of fold changes. Pathway analysis was done using the GeneMANIA Cytoscape plugin. Bar graphs were generated in Graphpad Prism. A) Proteins for which HT stimulus is

dominant over G stimulus. The theoretical expression patterns are depicted in the top panel (Red, upregulation; green, downregulation; and black no statistically significant change in expression). The heatmap of fold changes in expression for 30 proteins for which HT stimulus is dominant over G stimulus is shown in bottom panel. B) Proteins for which G stimulus is dominant over HT stimulus. The theoretical expression patterns are depicted in the top panel. The heatmap of fold changes in expressions for 121 proteins for which G stimulus is dominant over HT stimulus is shown in bottom panel. C) Bar graph shows the –log q-value of enrichments of the top five pathways in the list of proteins for which HT stimulus is dominant over G stimulus. D) Bar graph shows the –log q-value of enrichments of the top five pathways in the list of proteins for which G stimulus is dominant over HT stimulus.

For the environmental interactions in which the HT stimulus was dominant over the G stimulus, the p-values for all of the effect size estimates were less than 0.05. The changes for the HT and HT+G stimuli were in the same direction and differed from the G stimulus (Fig. 5A, top panel, rows 1 and 3). Alternatively, the p-values for only the HT and HT+G stimuli effect size estimates were less than 0.05 and the directions of change for the HT and HT+G stimuli were the same (Fig. 5A, top panel, rows 2 and 4). In all, we identified 30 proteins for which the response to the HT stimulus was dominant over the G stimulus (Fig. 5A and Table 1). We used pathway analysis to identify which protein classes were responding to the dominant stimulus. The group of proteins for which the HT stimulus was dominant included the heat shock response proteins HSP10, HSP60, SSA1, SSA2, and HSP150 (Fig. 5A bottom panel, and Table 1). Pathway analysis of these 30 proteins showed that the top five enriched pathways included protein folding, protein refolding, and unfolded protein binding (Fig. 5C, Table 5). These pathways are expected to be important for growth at higher temperatures (Richter, Haslbeck, and Buchner 2010; Riezman 2004; Åkerfelt, Morimoto, and Sistonen 2010; de Nadal, Ammerer, and Posas 2011).

For the environmental interaction in which the G stimulus is dominant, we saw a similar set of patterns as described above except the G stimulus dominates the HT stimulus (Fig. 5B, top panel). There are 121 proteins for which the response to the G stimulus was dominant over the HT stimulus (Fig. 5B, bottom panel and Table 1). The group of proteins for which the G stimulus was dominant includes metabolic enzymes such as CDC19, ACO1, and LSC1 (Fig. 5B, bottom panel, and Table 1). Pathway analysis of these 121 proteins showed that the top five pathways included the oxidation-reduction process, the generation of precursor metabolites and energy, and the tricarboxylic acid cycle (Fig. 5D, and Table 6). All of these three pathways are expected to be important for respiratory growth (Schüller 2003; Brisson et al. 2001; Nevoigt and Stahl 1997). Consistent with our hypothesis, pathway analysis of proteins that respond to a dominant environmental stimulus reveals a functional relationship to the response to the stimulus. High temperature has a dominant effect on proteins involved in protein folding, while glycerol has a dominant effect on proteins involved in respiratory metabolism. These results show the practical applications of using dominant environmental interactions to identify proteins that respond to specific stimuli and that are directly involved in the cell's response to that stimulus.

**Analysis of expression patterns reveals that environmental interactions mirror gene interactions.**

In addition to the dominant interactions of concurrent environmental stimuli, we observed other classes of environmental interactions that mirror gene interactions. First, we observed a class of proteins whose abundance either increased or decreased in response to both the individual stimuli as well as the concurrent stimuli (Fig. 6A). This

is similar to gene pairs in which both the individual mutants as well as the double mutant have the same phenotype. We classified these proteins as non-specific environmental responders. This class is represented by two theoretical expression patterns: activated or repressed (Fig. 6A, top panel and Table 1). For these non-specific environmental response modules, the p-values for all the effect size estimates were less than 0.05 and the directions of change were the same (Fig. 5A, top panel). We identified 175 proteins that correspond to these patterns, and pathway analysis revealed that they are largely involved in protein synthesis and translational control (Fig. 5A, bottom panel and 5D, and Table 7).

We also observed proteomic responses to concurrent environmental stimuli similar to gene interactions in which the two single mutants are wild-type or have one phenotype, while the double mutant has a different phenotype (Fig. 6B). This class includes proteins whose expression was either decreased or unchanged after a single stimulus but was increased if both stimuli were applied concurrently. The class also includes proteins whose expression was either increased or unchanged after a single stimulus but was decreased by the concurrent stimuli. We classified this environmental interaction group as a discordant class. There are eight theoretical expression profiles in the discordant environmental interaction class (Fig. 6B, top panel). For the discordant environmental interaction, the p-value for the HT+G concurrent stimuli effect size estimate was less than 0.05 and the directions of change for either the HT or G stimuli were not the same as HT+G. We identified 41 proteins that show discordance (Fig. 6B, bottom panel and Table 1). They are mainly involved in protein synthesis and metabolic pathways (Fig. 6E, and Table 8).

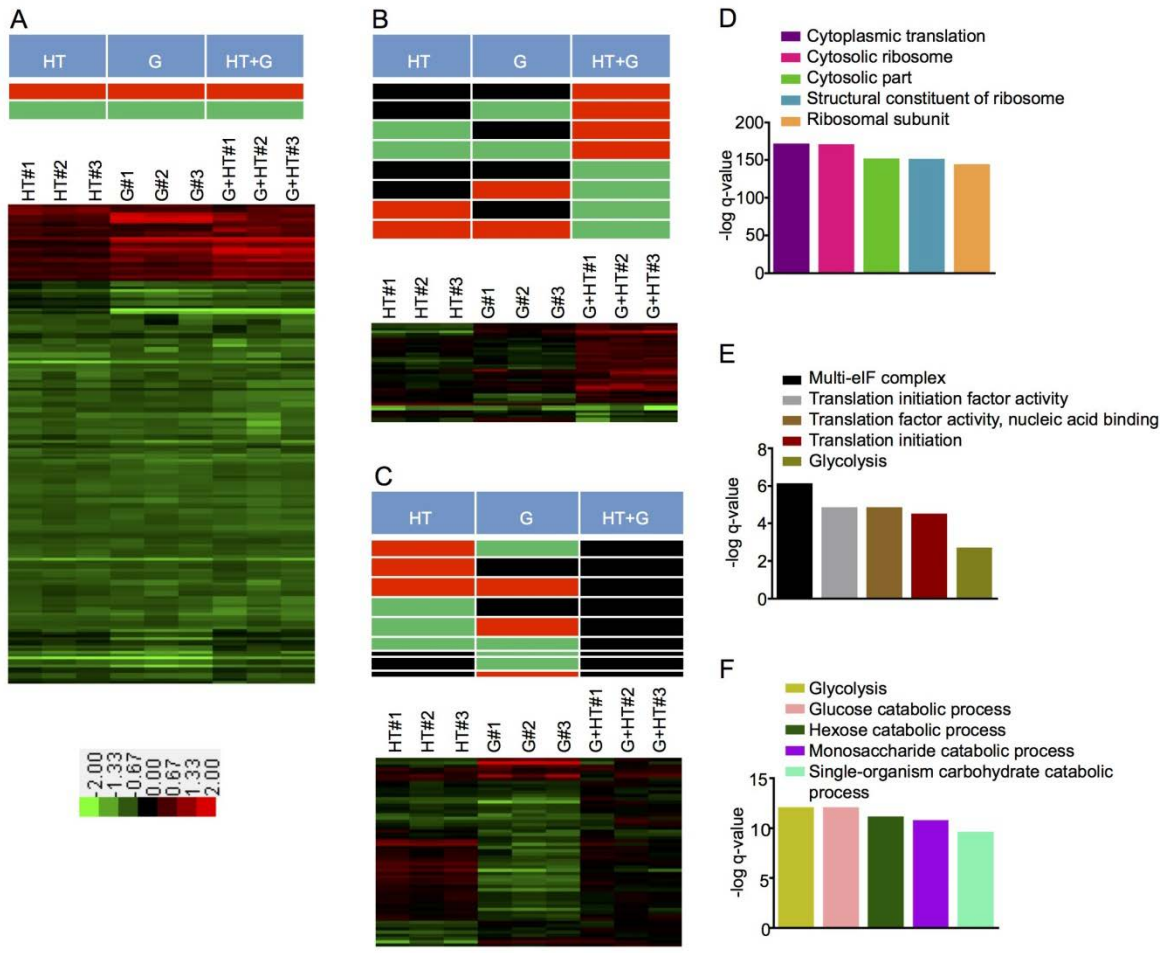



**Figure 6** - Proteins in different environmental interaction classes and the corresponding enriched pathways after concurrent G and HT stimuli

The color bar shows the range of fold changes. Pathway analysis was done using GeneMANIA Cytoscape plugin. Bar graphs were generated in Graphpad Prism. A) Non-specific environmental response (NER) proteins to individual and concurrent HT and G environmental stimuli. The theoretical expression patterns are shown in the top panel. The fold changes of 175 NER proteins are shown as a heatmap. B) The theoretical expression patterns for discordant environmental interaction are shown in the top panel. The fold changes of 41 proteins are shown as a heatmap. C) The theoretical expression patterns for suppression environmental interaction are shown in the top panel. The fold changes of the 58 proteins affected by suppression are shown as a heatmap. D) Bar graph shows the –log q-value of enrichments for the top 5 pathways for the non-specific environmental response proteins. E) Bar graph

shows the –log q-value of enrichments for the top 5 pathways in the list of proteins affected by discordant environmental interaction. F) Bar graph shows the –log q-value of enrichments for the top 5 pathways in the list of proteins affected by suppression environmental interaction.

Finally, we observed suppression, in which a protein's abundance changed in response to a single stimulus, yet the change was suppressed by the simultaneous application of the second stimulus (Fig. 6C). This class is similar to gene interactions in which double mutants show the wild-type phenotype (Dixon et al. 2009; St Johnston 2002). The suppression class is represented by eight theoretical expression patterns (Fig. 3C, top panel). For suppression environmental interactions, the p-value for the HT+G effect size estimate was more than 0.05, and the p-value for at least one of HT and G stimuli effect size estimates was less than 0.05. We identified 58 proteins that are affected by suppression (Fig. 6C, bottom panel and Table 1). Pathway analysis revealed that metabolic pathways are most affected by suppression (Fig. 6F and Table 9).

**A large fraction of the proteome is affected by environmental epistasis.**

An important feature of genetic epistasis is that the modulating effects of multiple genes are not always independent of each other (Cordell 2002; P. C. Phillips 2008; Fisher 1958; Visser, Cooper, and Elena 2011; Mani et al. 2008). In many cases, non-independence is diagnostic of a functional relationship between genes (Cordell 2002; P. C. Phillips 1998; Visser, Cooper, and Elena 2011). Genetic epistasis is used to test if the effects of genetic elements are independent. Genetic epistasis occurs when the effects are not independent. We tested if the effects of these two individual environmental stimuli were independent of each other for individual proteins in the

proteome. Similar to the mathematical approach to genetic epistasis, we measured the response of each protein and classified a response as influenced by environmental epistasis if the sum of the effects of the individual stimuli for a protein was not equal to the response to the concurrent stimuli (t-test, p-value ≤0.05) (Cordell 2002; Fisher 1958; P. C. Phillips 2008). We used log2 fold change as the measure of the effect of a stimulus. From our list of 466 quantitated proteins, 240 proteins were affected by environmental epistasis (Table 1). Pathway analysis of these proteins revealed that a majority of the enriched pathways are involved in protein synthesis and translation control (Fig. 7A and Table 10). The topmost enriched pathways included cytoplasmic translation, cytosolic ribosome, and structural constituent of ribosome (Fig. 7A and Table 11).

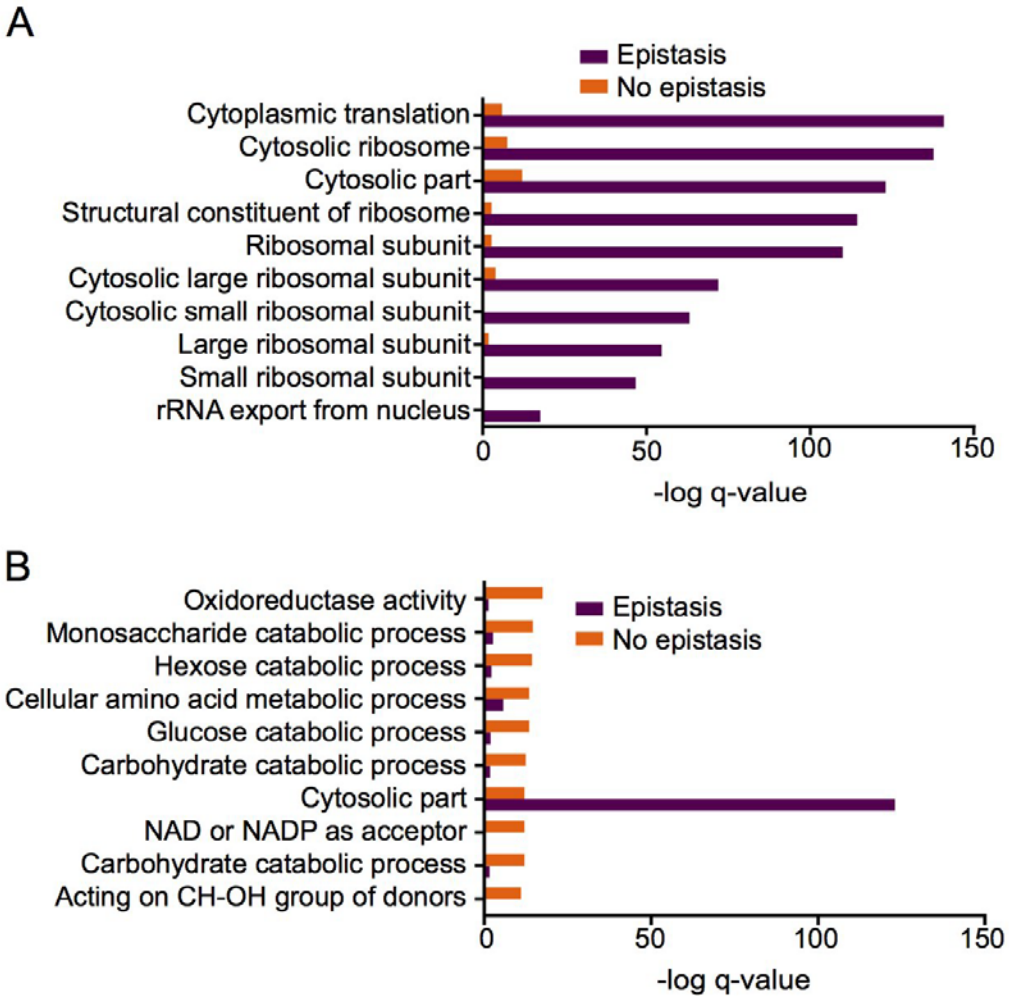



**Figure 7** - Environmental epistasis in the proteomic response to concurrent stimuli

Pathway analysis was done using the GeneMANIA Cytoscape plugin. Bar graphs were generated in Graphpad Prism. A) Bar graph shows the –log q-value of enrichments of the top 10 pathways in the list of proteins affected by epistasis (purple) and their –log q-value in the list of proteins not affected by epistasis (orange). B) Bar graph shows the–log q-value of enrichments of the top 10 pathways in the list of proteins not affected by epistasis (orange) and their –log q-value in the list of proteins affected by epistasis (purple).

Pathway analysis of the 226 proteins not affected by environmental epistasis revealed a large number of metabolic pathways (Fig. 7B and Tables 1 and 10). It is interesting to note that the distribution of pathways affected by environmental epistasis is different from those that are unaffected. Protein synthesis and translational control seems to be disproportionately affected by environmental epistasis compared to other pathways. These pathways have previously been found to change in response to the changes in the growth rate (Regenberg et al. 2006; Slavov and Botstein 2011). If the effects of the two stimuli on the growth rate are not independent, it could explain the observed environmental epistasis. To test the independence in the effects of the two stimuli on the growth rate, we determined the doubling times under the same conditions. The change in the doubling times was used to measure the effect of a stimulus. Our data shows that the effects of high temperature and glycerol on the growth rate are additive and, therefore, independent of each other (Fig. 8). Further studies are required to elucidate the functional significance of the environmental epistasis.

A number of genetic epistasis subtypes have been defined based upon the mathematical models used to measure the expectation of a phenotype in double mutants (Visser, Cooper, and Elena 2011; Mani et al. 2008; Gao, Granka, and Feldman 2010; Hallgrímsdóttir and Yuster 2008; Li and Reich 2000). Four most commonly used definitions are (1) additive, (2) multiplicative, (3) minimum, and (4) log (Mani et al. 2008; Gao, Granka, and Feldman 2010). Although we used only the additive definition for developing the concept of the environmental epistasis in this study, future studies can be performed to compare the results obtained using different definitions.

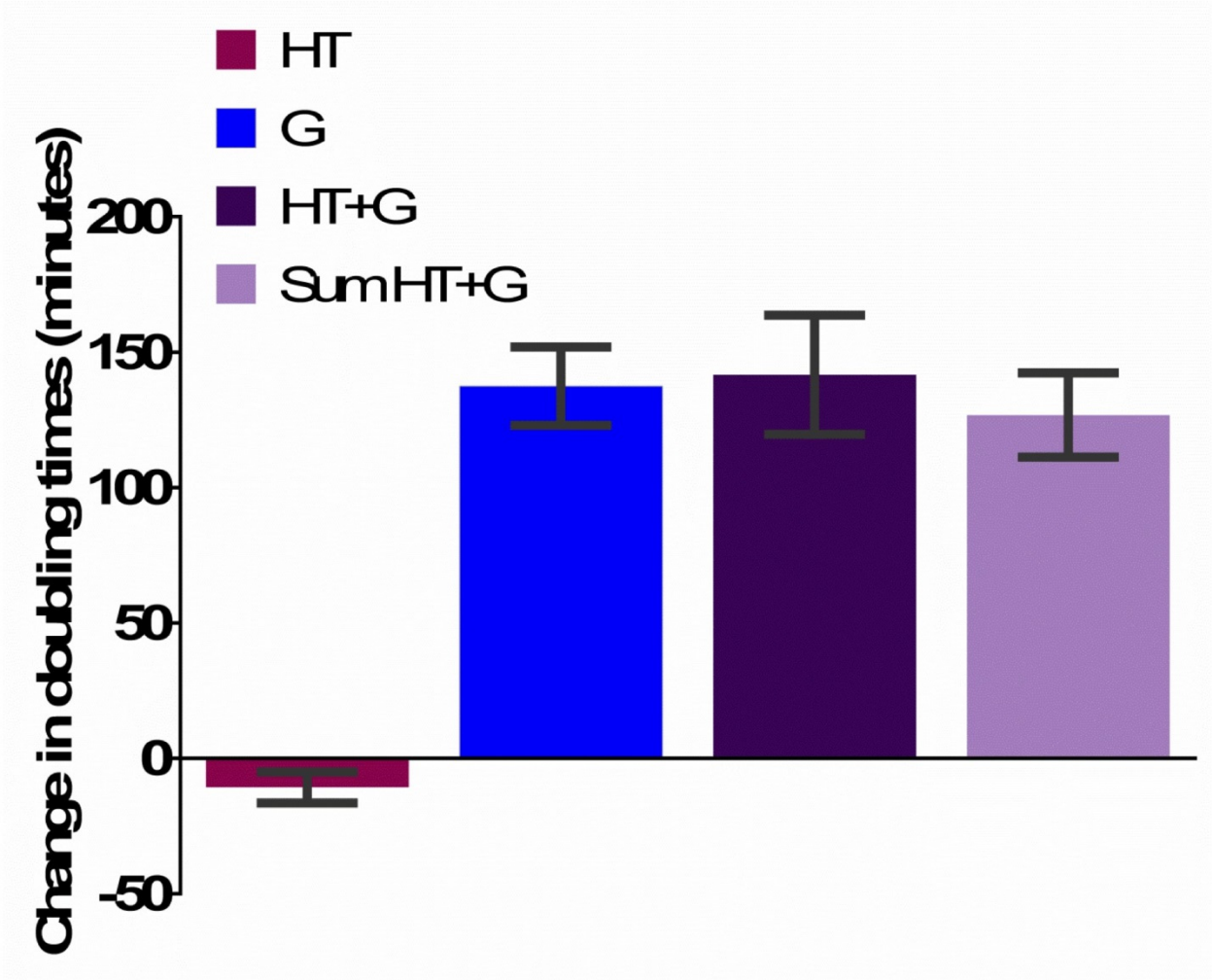



**Figure 8 -** The effect of high temperature and glycerol on yeast doubling times

Doubling times were calculated for growth in control (n=25), high temperature (n=25), glycerol (n=25), and concurrent high temperature and glycerol (n=24). The difference in doubling times from the control was used to measure the effect of the stimuli and is plotted on Y-axis. HT leads to a decrease of -11 minutes (sd = 6), G leads to an increase of 137 minutes (sd = 14), and HT+G leads to an increase of 142. minutes (sd = 22). The expected effect of HT+G was calculated by summing the observed effects of HT and G (Sum HT+G, increase of 127 minutes with sd of 16). The difference in the means for HT+G and Sum HT+G was not statistically significant (p-value = 0.1034, two-tailed t-test of independence with Bonferroni correction for 11 comparisons)

**Environmental interactions and epistasis regulate mRNA levels.**

Although, we identified the environmental interactions using quantitative proteomic data, we speculated that this framework would be applicable to any quantifiable readout including transcriptomic and phenotypic traits. In pioneering studies using chemostat cultures of S. cerevisiae, Knijnenburg et al. measured the transcriptional response of yeast to multiple, concurrent environmental stimuli (Knijnenburg et al. 2009). They found linear regression models of expression for the vast majority of genes required a combinatorial interaction term (Knijnenburg et al. 2009). This suggests the change in transcription of most genes cannot be explained by simply adding the effects of the individual stimuli. Based on our proteomic results, we hypothesized that environmental epistasis plays a role in determining the state of the transcriptome as well.

To test if our environmental interaction and epistasis models are observed in the transcriptomic responses to concurrent stimuli, we analyzed Knijnenburg dataset which measured the transcriptomic responses of yeast cells growing in carbon limited chemostat cultures (Knijnenburg et al. 2009). In the experiment, two concurrent stimuli were applied: (1) a change in nitrogen source from ammonium sulfate to methionine and (2) a change from aerobic to anaerobic growth (Fig. 9A and Table 12) (Knijnenburg et al. 2009). The data showed 564 transcripts were affected by environmental epistasis, while 5987 transcripts were not affected (p-value ≤0.05) (Table 12). In contrast to our proteomic analysis, pathway analysis of the transcripts affected by environmental epistasis revealed enrichment for pathways including microbody, peroxisome, and phytosteroid metabolic process (Table 13). This could be because of the differences

between the strains, stimuli, and culture conditions used in the transcriptomic and our proteomic studies. Similar to our proteomic analysis, we observed dominant environmental interactions in the expression of the transcripts (Fig. 9B and 9C and Table 12). Nitrogen source was dominant for 281 transcripts (Fig. 9B and Table 12). Pathway analysis of these transcripts identified pathways involved in methionine metabolism such as sulfur amino acid metabolic process, sulfur compound metabolic process and methionine metabolic process (Fig. 9D and Table 14). Similarly, anaerobic growth was dominant for 938 transcripts (Fig. 9C and Table 12). Pathway analysis of these differentially expressed transcripts showed enrichment of pathways involved in energy production such as cellular respiration, mitochondrial membrane and respiratory chain (Fig. 9E and Table 15). We also observed the same environmental interaction classes in their transcriptomic data as in our proteomic data, including non-specific environmental response, discordance, and suppression (Table 12). These results strongly suggest that environmental interactions play a significant role in regulating the biochemical state of cells.

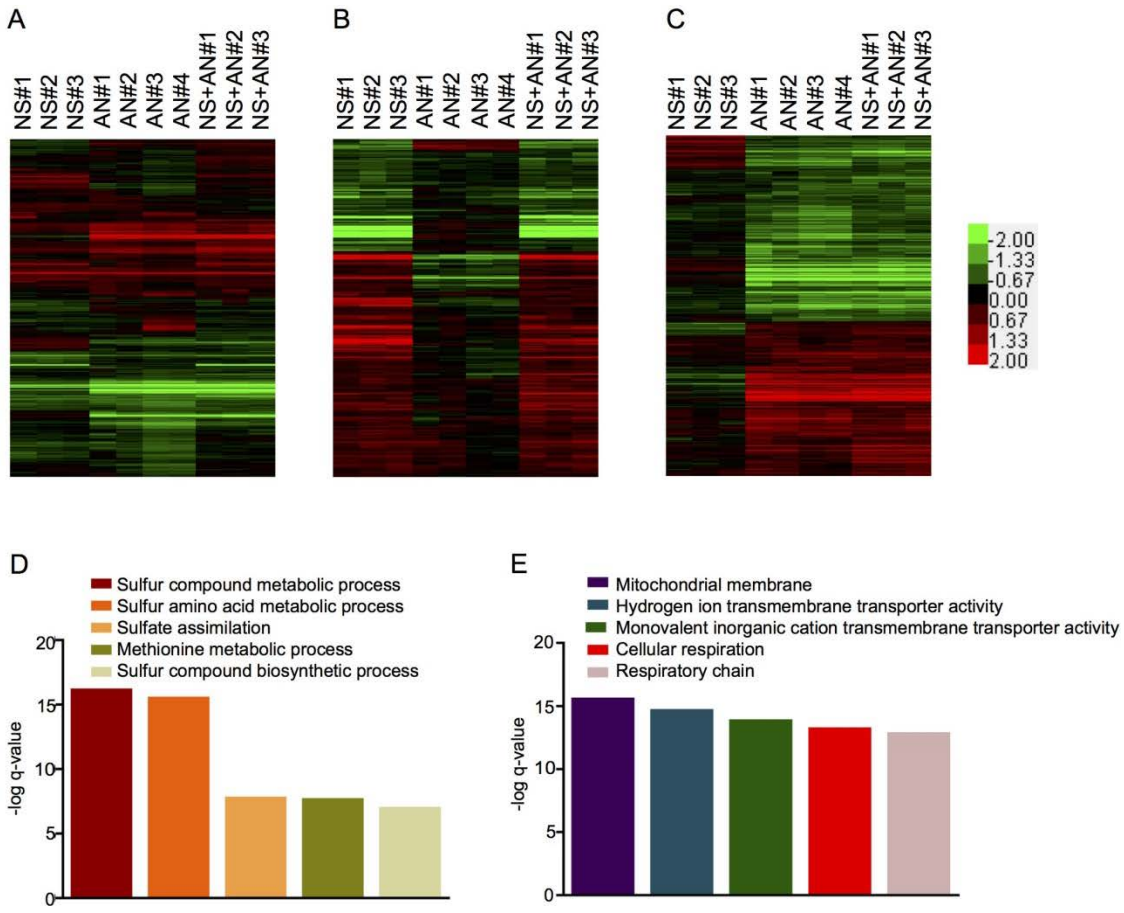



**Figure 9 -** Environmental interactions affect transcriptomic profiles as well

Normalized expression data from Knijnenburg et. al. 2009 was used for the analyses. The transcriptomic data used in the study used haploid S. cerevisiae cells (CEN.PK113-7D MATa) grown in carbon limited chemostat cultures under 4 conditions – 1) ammonium sulfate as nitrogen source (n=5), 2) methionine as nitrogen source, NS stimulus (n=3), 3) Anaerobic condition, AN stimulus (n=4), and 4) methionine as nitrogen source under anaerobic conditions NS+AN stimulus (n=3) (13). Fold changes were calculated from normalized expression data using average normalized expression data from the five replicates of growth with ammonium sulfate as the baseline. The color bar shows the range of fold changes. Pathway analysis was done using GeneMANIA Cytoscape plugin. Bar graphs were generated in Graphpad Prism. A) A heatmap of fold changes of the complete transcriptomics dataset consisting of 6551 transcripts. B) A heatmap showing the fold changes for 281 transcripts for which NS stimulus is dominant. C) The –log q-value of enrichment for the top 5 pathways enriched in the list of transcripts for which NS stimulus is

dominant. As anticipated, pathways expected to be involved in metabolization of methionine are enriched. D) A heatmap showing the fold changes for 938 transcripts for which AN stimulus is dominant. E) The –log q-value of enrichment for the top 5 pathways enriched in the list of transcripts for which AN stimulus is dominant. As anticipated, pathways expected to be involved in energy production are enriched.

## Coexpression network analysis shows community structures are guided by environmental interaction and epistasis.

Coexpression networks link together proteins whose expression levels are regulated in the same way (Stuart et al. 2003; B. Zhang and Horvath 2005). As a consequence, coexpression network analysis can be used to determine if the abundances of proteins affected by environmental epistasis are regulated differently than the proteins that are not affected by environmental epistasis. To explore the protein modules whose expression changes are correlated with each other, we built a coexpression network using the merged proteomic responses from both individual and concurrent stimuli using the Sparse PArtial Correlation Estimation approach (SPACE) (Fig. 2A) (Peng et al. 2009). An edge, representing coexpression, was introduced between two proteins if the correlation between them was above the average of the correlation matrix. To validate the network, we first tested the power law structure of the reconstructed network (Peng et al. 2009; Clauset, Shalizi, and Newman 2009). The reconstructed network followed the power law distribution. The power law parameter α was approximately 4, which is close to the empirically observed value of 3.45 (Clauset, Shalizi, and Newman 2009).

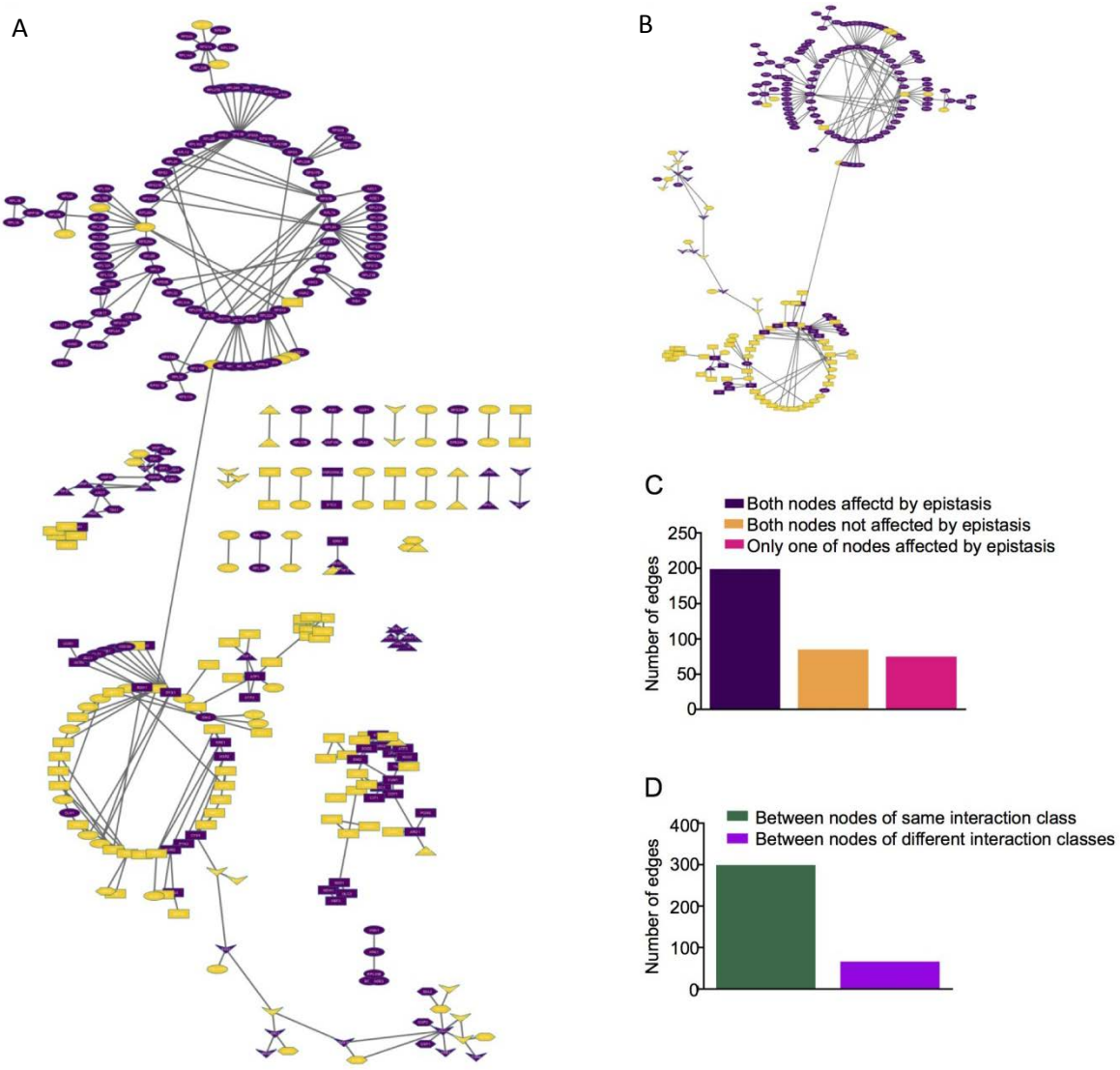



**Figure 10 - Coexpression network based on all the quantified proteins and all conditions**

Proteins are depicted as nodes. Nodes that are coexpressed are connected with an edge. The coexpression network was generated with SPACE algorithm using fold change. Network visualization and analysis was done in Cytoscape 3.1.1. Bar graphs were generated in Graphpad prism. A) All nodes that have at least one edge. Nodes affected by environmental epistasis are highlighted in purple. The circular layout was used to generate the initial network graphics in Cytoscape 3.1.1. Far-flung communities of inter-connected nodes were manually brought together, while preserving the inner community structure, for better visualization. B) The largest community in the coexpression network. Most of the proteins affected by environmental epistasis are members of a subgraph (top circle) that predominantly contains

other proteins that are also affected by environmental epistasis. A similar trend is observed with the proteins not affected by environmental epistasis (bottom circle). C) The numbers of three types edges: 1) both nodes are affected by environmental epistasis (199 edges), 2) neither of the nodes are affected by environmental epistasis (85 edges), and 3) only one of the nodes is affected by environmental epistasis (75 edges). Proteins affected by epistasis are predominantly connected to proteins that are also affected by epistasis. D) Number of edges that connect nodes to other nodes within the same environmental interaction classes (299) or between the classes (60). Co-regulatory connections between proteins are predominantly between those of the same class.

Next, we repeatedly reconstructed the network by varying the tuning parameter around the default value and fitting the network to the power law distribution. We found that the reconstructed network follows the power law distribution and that the power law parameter was in the range of 3.75. We identified the sub-graph spanned by the top 1% of highly connected nodes. We found that the Jaccard similarity score of these highly connected nodes was 0.83 on the scale of 0 to 1. Therefore, these nodes were classified as hub nodes, which is one of the characteristic features of power law networks. There were 7 hub nodes based upon the above criterion. Next, we checked the significance of the identified hubs using the Wilcox Rank sum test and found that the hub community is statistically significant (p-value = 0.04) (Kolaczyk and Csárdi 2014). Finally, we compared the reconstructed network with BioGrid protein interaction data and found that approximately 30% of the edges are previously known interactions and that these interactions were found in every reconstructed network when we varied the tuning parameter to estimate the partial correlation matrix (Stark et al. 2006). The final coexpression network consisted of 329 nodes with at least one neighbor and a total of 359 edges (Fig. 10A).

The largest community within this network includes 205 nodes and 249 edges, with two clearly separate sub-graphs connected by a single node (Fig. 10B). Interestingly, one sub-graph consists predominantly of proteins affected by environmental epistasis while the second sub-graph consists of proteins not affected by environmental epistasis. Within the global coexpression network, we observed that proteins affected by epistasis were more likely to be linked with each other than with proteins that are not affected by epistasis and vice versa (Fig 10A). There are 199 edges between two proteins affected by epistasis and 85 edges between two proteins not affected by epistasis. However, only 75 edges involved proteins of both types (Fig. 10C). This structural organization of the coexpression network suggests that the responses of proteins affected by environmental epistasis are controlled by a different mechanism than the responses of those not affected by environmental epistasis.

Previous studies indicate that proteins linked in a coexpression network are likely to function in the same pathway (Stuart et al. 2003). We hypothesized that the grouping of proteins upon classification into environmental interaction classes might be driven by their functional associations. If true, we would expect to find more edges in the coexpression network between proteins within the environmental classes. Indeed, we found this result in this network. Our data show that 299 of the edges (83%) are between proteins in the same environmental interaction class, while only 60 are between proteins in different classes (Fig 10D).

**Discussion**

Using the concepts of gene interactions and epistasis, we have developed a unifying conceptual framework to understand the cellular responses to complex

environmental stimuli. Although, we have only explored the cases with complete dominance of a stimulus, it is possible that both the stimuli contribute to a change in expression. It is also possible that many stimuli contribute towards a change. We speculate that the tools and approaches developed for gene-gene interactions involving multiple genes can be applied in such cases (Cordell 2009). In addition to linear regression modeling and ANOVA, we also tested our hypothesis using one sample and two sample t-tests of independence. The results from both approaches were in good agreement.

The effect of mixtures of compounds has been actively studied in toxicology, especially in the context of environmental toxins (Altenburger et al. 2013; Hermens, Leeuwangh, and Musch 1985; Belden, Gilliom, and Lydy 2007; Altenburger, Nendza, and Schüürmann 2003; Altenburger et al. 2012; Altenburger, Walter, and Grote 2004; Berenbaum 1989; Deneer 2000; Greco, Bravo, and Parsons 1995; J.-H. Lee and Landrum 2006; Schoen 1996; Faust et al. 2001). These studies have led to the development of three complementary models to predict the combined effects of compounds in a mixture: (1) in the concentration addition model the total toxicity of a mixture is the sum of the individual toxicities of the component compounds, (2) in the independent action model the toxicities of the components of a mixture are independent of each other, and (3) in the simple interaction model the individual components, at the concentrations being tested, are not toxic, but are toxic when used together in a mixture. These models have been successful in predicting the total toxic effects of mixtures of compounds in many cases (Hermens, Leeuwangh, and Musch 1985; Belden, Gilliom, and Lydy 2007; Altenburger, Nendza, and Schüürmann 2003;

Altenburger et al. 2012; Altenburger, Walter, and Grote 2004; Deneer 2000; Faust et al. 2001) . However, it is not immediately clear which one to apply in a specific case without model fitting (Belden, Gilliom, and Lydy 2007).

Environmental interactions and epistasis can be extrapolated to explain the three models. For example, the concentration addition model can be the case of incomplete dominance where many stimuli affect the biological processes under investigation. This would happen if the compounds in the mixture affect similar biological pathways. If the actions of the compounds are antagonistic to each other, it may lead to either the dominance or the suppression interaction. If their actions are not antagonistic, the combined effect would be the sum of the individual effects which could be observed as the non-specific environmental response.

The independent action model explains the case where the compounds under investigation act upon different pathways (Altenburger et al. 2013; Altenburger, Nendza, and Schüürmann 2003; Altenburger et al. 2012; Altenburger, Walter, and Grote 2004; Greco, Bravo, and Parsons 1995; Schoen 1996). This is similar to a gene interaction where two mutations have two unrelated phenotypes and both phenotypes persist in the double mutant. By applying the logic of environmental interaction to this model, we can deduce that the changes induced by a mixture that follows the independent action model would have elements specific to the component compounds of the mixture. Additionally, the changes important to a specific compound would persist in the combinatorial condition, which could be used to identify molecules and pathways that respond to the specific compound in the mixture.

The simple interaction model explains the cases where the compounds individually have little or no toxicity, but are toxic when applied together (Berenbaum 1989; Greco, Bravo, and Parsons 1995). In terms of environmental interaction, this could be a case of the discordance interaction. The effects explained by this model could also be a special case of environmental epistasis, where the combined effect of compounds is more than the sum of their individual effects. It is worth noting that although we discuss only three of the mixture toxicity models, there are a number of other models that explain the toxicities of compounds in a mixture (Hermens, Leeuwangh, and Musch 1985; Altenburger et al. 2013; Belden, Gilliom, and Lydy 2007; Altenburger, Nendza, and Schüürmann 2003; Altenburger et al. 2012; Altenburger, Walter, and Grote 2004; Berenbaum 1989; Deneer 2000; Greco, Bravo, and Parsons 1995; J.-H. Lee and Landrum 2006; Schoen 1996; Faust et al. 2001). Environmental interactions and epistasis provides a conceptual framework unifying the different toxicity models. The interpretation of results can be made simpler using environmental interactions and epistasis.

Phenotypic plasticity provides the conceptual framework for studying the interaction between genotype and environment. Phenotypic plasticity is the ability of an organism to change its phenotype in response to changes in the environment (Scheiner 1993). It has been used to explain the ability of the same genotype to generate different phenotypes in different environments (Scheiner 1993). However, phenotypic plasticity considers the environment as a monolithic entity. It fails to separate the relative contributions of the different environment components, for example; physical components such as temperature and pressure, chemical components such as

nutrients, and signaling molecules that activate different pathways. Applying environmental interactions and epistasis would help parse out the individual contributions of the stimuli towards the change in the phenotype.

Similar to genetic epistasis, our data show that the effects of individual environmental stimuli are not necessarily additive. Proteins affected by environmental epistasis are distributed throughout the genome and do not appear to be clustered at specific locations in the genome (Fig. 11). The prevalence of environmental epistasis in determining the changes in the proteome suggests that epistasis needs to be taken into account when building mathematical models of gene expression.

Consideration of environmental epistasis is especially important in light of the recent attempts to build quantitative linear regression models of gene expression in which the independent variables are the environmental stimuli and the dependent variable is gene expression (Nagano et al. 2012). Interestingly, in a linear regression modeling study of transcriptional regulation in rice under native conditions, the regression model was able to predict gene expression under native conditions even if the environmental parameters varied slightly from those used for building the model. However, the predictive power of the regression model was reduced under controlled laboratory conditions suggesting that there may have been unknown epistatic interactions in the native conditions absent in the controlled lab conditions (Nagano et al. 2012).

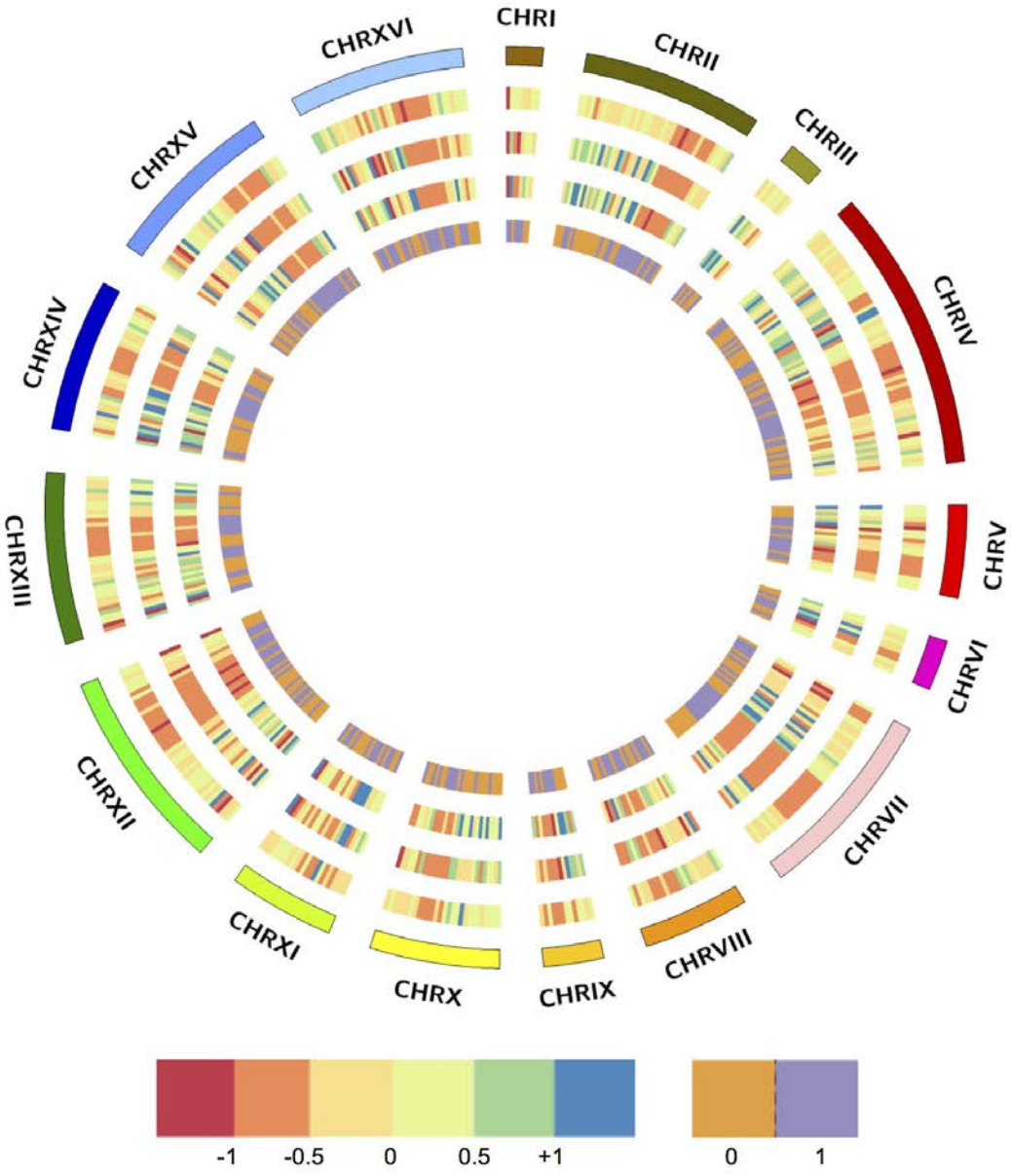



**Figure 11 -** Visualization of S. cerevisiae genomic locations of the proteins quantitated with fold changes represented as a heatmap using Circos plot

Outermost circle- chromosomes, Second circle-fold changes of proteins with HT stimulus, Third circle-fold changes of proteins with G stimulus, Fourth circle-fold changes of proteins with HT+G stimuli, innermost circle-whether affected by epistasis or not (Purple: Affected by environmental epistasis, Orange: Not affected by environmental epistasis).

Concurrently applied environmental stimuli behave similarly to genetic elements in the way they interact to regulate the biochemical states of the cells. The observation of environmental interactions and epistasis in determining the states of both the proteome and transcriptome in diverse experimental conditions suggests the prevalence of this phenomenon in nature. Essentially, environmental interaction in concert with phenotypic plasticity and gene interactions can be envisaged as a mathematical operator with three components that determines the changes in the biochemical state of the cell. The gene interaction component is derived from the effects of the genetic elements, while the environmental interaction component results from the effects of all the environmental stimuli. When the gene and environmental interactions are not independent of each other, phenotypic plasticity accounts for the deviations of the observed from the expected characteristic or trait. Most studies so far have treated phenotypic plasticity, gene interactions, and environmental interactions separately due to a lack of a common unifying framework (Cordell 2002; P. C. Phillips 1998; St Johnston 2002; Visser, Cooper, and Elena 2011; Mani et al. 2008; Altenburger et al. 2013; Belden, Gilliom, and Lydy 2007; Altenburger, Nendza, and Schüürmann 2003; Altenburger, Walter, and Grote 2004; Greco, Bravo, and Parsons 1995; Scheiner 1993; Via and Lande 1985; Carl D. Schlichting and Levin 1984; Gerard, Vancassel, and Laffort 1993; C. D. Schlichting and Pigliucci 1993; Wilson and Lindow 1993; Tonsor, Elnaccash, and Scheiner 2013). Our data suggest that as an abstraction, environmental stimuli can be treated as genes to build a conceptual framework that combines the effects of genes and stimuli. Environmental interactions and epistasis play a critical role

in cellular homeostasis as seen in this study's patterns of change in the proteome and the transcriptome.

Our data also suggest that a protein or a transcript is more likely to be critical for responding to a dominant environmental stimulus than to a recessive one. This could lead to more efficient experiment designs for identifying factors directly affected by an environmental stimulus. For example, experiments could be designed in which an unrelated stimulus B is applied concurrently with the stimulus of interest A. The proteins or transcripts, for which the effect of A is dominant, would be more likely to be directly affected by stimulus A. We speculate that the same approach may be extended to genetic perturbations. In this case, an environmental stimulus could be applied in conjunction with the genetic perturbation. As with two concurrent environmental stimuli, a transcript or a protein for which the genetic perturbation is dominant may be more likely to be directly affected by it. Therefore, using dominance, environmental interactions can also be used to devise studies to identify agents, such as regulatory RNAs, proteins, or small molecules which are critical for driving a range of biological processes in health and disease including drug interactions, adaptation in tumor microenvironment and immune responses.

# Chapter III

## Carbon Source Alters the Protein Composition of Ribosomes for Translational Control


### Abstract

Ribosomes are the catalytic engines that drive protein synthesis. More recently, the role of ribosome in gene expression regulation has come under increased focus. In this study, I have explored the role of ribosomal proteins in translational control. I used iTRAQ labeling followed by liquid chromatography tandem mass spectrometry based protein quantitation and cryo-electron microscopy single particle classification and reconstruction to assay the changes in the protein composition of the ribosomes. I identified Rpl8a and Rpl8b as a candidate paralog pair whose change in abundance on the ribosomes is important for the ribosome filter mediated translational control. My data using yeast genetics and polysome profiling shows that Rpl8a and Rpl8b are not completely interchangeable. I found evidence supporting the presence of substoichiometric ribosomes and regulation of their proportions in response to changes in the carbon source.


### Introduction

Gene expression can be regulated at multiple levels, including transcription and translation. Regulation of translation, also known as translational control, is a major mechanism modulating eukaryotic gene expression. Translation is the process, driven by ribosomes as the catalytic engines, by which the information encoded in an mRNA is used to synthesize a protein. Translational control is the regulatory mechanism through

which the protein synthesis machinery is regulated to control the information flow from mRNA transcripts to proteins. The protein and RNA accessory factors play an active role in this process. The ribosomes were initially considered passive players (Kapp and Lorsch 2004; Dever and Green 2012; Sonenberg and Hinnebusch 2009; Hinnebusch 2015).

In recent years, the idea that the ribosomes are regulatory elements in gene expression regulation has been gaining ground. The role of ribosome in the regulation of gene expression is a very active area of research (Ruggero and Pandolfi 2003; Kondrashov et al. 2011; Xue et al. 2015; Jonathan R. Warner and McIntosh 2009; Jonathan R. Warner 2015; McIntosh and Warner 2007; Komili et al. 2007). Two complementary mechanisms have been proposed to explain the regulatory functions of ribosomes. In the first model, ribosomes act as reservoirs of regulatory molecules that are released upon specific cellular cues. This model is called the depot hypothesis (Ray, Arif, and Fox 2007; Mazumder et al. 2003). An important example of this mode of action is the role of RPL13A protein in inflammatory response (Mazumder et al. 2003; Kapasi et al. 2007).

The second model is based upon the heterogeneity in the protein and rRNA compositions of the ribosomes. In baker's yeast *Saccharomyces cerevisiae*, 138 ribosomal protein genes are present. This includes 59 duplicated paralog pairs. The amino acid sequences of the paralogs are not identical. There are nearly 150 copies of rRNA genes in yeast. These copies are not identical (McIntosh and Warner 2007; Komili et al. 2007; Jonathan R Warner 1999). Ribosomal proteins have been found to be post-translationally modified. Post-translational modifications add another level of complexity

to the ribosomal subpopulations (Martin et al. 2014; S. Ramagopal 1992). Taking these observations into account, Vincent Mauro and Gerald Edelman proposed the ribosome filter hypothesis. According to the hypothesis, different ribosomes with differing composition translate specific mRNAs more efficiently (Mauro and Edelman 2002; Vincent P Mauro and Gerald M Edelman 2007).

In addition to the heterogeneity in ribosome composition due to the use of a specific paralog, a post-translational modification and a difference in the rRNA sequence, two other potential sources of heterogeneity involving the core components of the ribosomes are possible. In one model, a ribosomal protein can be present in more than one copy on the ribosome. We call this the superstoichiometric composition model in which subpopulations of ribosomes carrying extra copies of a ribosomal protein can act either as a depot of the regulatory extra copy ribosomal protein, or translate certain mRNA with higher efficiency. In the second model, a ribosome can be missing ribosomal protein(s) or sequences of rRNA. We call this the substoichiometric model. Similar to the superstoichiometric model, these ribosomes might have acted as a depot of the regulatory missing ribosomal protein(s). Alternatively, the substoichiometric ribosome can translate specific mRNAs at a higher efficiency.

The publication of ribosome filter hypothesis has provided a conceptual framework in which to understand the role of ribosomal heterogeneity, and the ribosome has received increased attention as a regulatory factor in recent years (Xue and Barna 2012; McIntosh and Warner 2007). The ribosome filter hypothesis has gained support from many subsequent studies (A. S.-Y. Lee, Burdeinick-Kerr, and Whelan 2013; Kondrashov et al. 2011; Komili et al. 2007; Xue et al. 2015). A large body of work has

concentrated on the rRNAs and suggests that ribosomes are indeed regulatory elements, and that rRNA heterogeneity arising from multiple copies of rRNA genes plays an important role in this process (Vincent P Mauro and Gerald M Edelman 2007; Mauro and Edelman 1997; Owens et al. 2001; Hu et al. 1999).

There is also evidence that suggests the heterogeneity in ribosomal proteins results in differences in translational efficiency (S. Ramagopal 1992). A comparison of protein composition of ribosomes from skeletal muscles and liver in rats using 2-dimensional gel electrophoresis revealed differences between the ribosomes from the two tissues (Sherton and Wool 1974). This suggested a differential requirement of ribosomal proteins in different mammalian tissues. *Dictyostelium discoideum*, ribosomes from spores and vegetative cells differ in protein composition and posttranslational modification (S. Ramagopal 1992). This suggested a differential temporal requirement of ribosomal proteins and their post-translation modifications in organismal development. In S. cerevisiae, deletion of one of the paralog pairs often results in a different phenotype from that of the other paralog, indicating that that the different paralogs have different roles (Komili et al. 2007; Giaever et al. 2002; Breslow et al. 2008). Pamela Silver and co-workers demonstrated that translation of the S. cerevisiae ASH1 mRNA is more efficient in the presence of particular paralogs (Komili et al. 2007). Studies in Maria Barna lab has shown that Rpl38 is needed for efficient translation of specific *Hox* mRNAs (Xue et al. 2015; Kondrashov et al. 2011). In another study, Rpl40 was found to be required for translation initiation of vesicular stomatitis virus (VSV) mRNAs (A. S.-Y. Lee, Burdeinick-Kerr, and Whelan 2013). Hyper-phosphorylation of S6 has been implicated in upregulation of protein synthesis (Thomas et al. 1982; Duncan

and McCONKEY 1982). The phosphorylation of S6 has been proposed to increase the affinity of the ribosomes for TOP-element containing mRNAs, although there are conflicting observations in mice (Ruvinsky et al. 2005; Volarević and Thomas 2000; Ruvinsky and Meyuhas 2006).

Taken together, these studies have provided strong evidence in support of the ribosome filter hypothesis. However, the hypothesis is yet to be directly tested. This has been mainly because of the difficulty in identifying both the ribosomes with a specific composition and the mRNAs that they translate more efficiently. To identify both the ribosome and the transcripts, I devised a simple strategy based upon a corollary of the ribosome filter hypothesis. Cells growing in one growth condition require a specific proteome that is optimum for that condition. Cells growing in a different condition will require a different proteome. Therefore, if the ribosome filter hypothesis is correct, the complement of ribosomes required to synthesize the two proteomes will be different (Figure 1). Quantification of the ribosomal proteins in the purified ribosomes should allow identification of the paralogs whose requirements are different. Once a paralog has been identified, ribosome profiling of null mutants will allow identification of the transcripts whose translation is affected by specific paralogs. Using this strategy in yeast cells growing with glucose or glycerol as carbon source, I have identified the paralog pair Rpl8a and Rpl8b as candidate ribosomal proteins with differential requirements for specific transcripts. Polysome profile analysis using heterozygous diploid null mutants, *rpl8a* and *rpl8b*, suggests that the functions of Rpl8ap and rpl8bp are not inter-changeable. Using cryo-electron microscopy in collaboration with the Joachim Frank lab, I have found evidence in favor of substoichiometric ribosomes. A

time course experiment coupled with cryo-EM revealed that the proportions of substoichiometric ribosomes changed in response to change in carbon source from glucose to glycerol.

**Materials and methods**

**Strains and Media**. All yeast media, growth, and genetic manipulation was done using standard techniques (Amberg, Burke, and Strathern 2005).The diploid strain BY4743 has been previously described (Baker Brachmann et al. 1998).

**Preparation of protein extract and ribosome purification**: Three biological replicates were used for each growth conditions tested, YPD at 30 $^{o}$C and YPG at 30 $^{o}$C. A 5ml overnight culture in YPD (1% yeast extract, 2% peptone, 2 % glucose) was inoculated from a single yeast colony from a YPD agar plate. Fifty ml of YPD at 30 $^{o}$C was inoculated with 50 µL of the overnight culture. One hundred ml of YPG at 30 $^{o}$C (1% yeast extract, 2% peptone, 3% glycerol) inoculated with 1 ml of the overnight culture. The twelve cultures were grown side-by-side with constant shaking at 175 rpm in an Innova 44 shaker incubator (New Brunswick Scientific). For all four growth conditions, cells were harvested at mid-log phase as determined by OD600 measurements. Cells growing in YPD were harvested after 14 h, while cells in YPG were harvested after 24 h. All cultures were centrifuged at 2000 rpm for 5 min at 4 $^{o}$C using a Sorvall HLR6/H600A/HBB6 rotor in Sorvall RC-3B centrifuge and washed with ice cold deionized H20. The cell pellets were resuspended in 1 mL ice cold wash buffer (10 mM Tris pH 8, 5 mM beta- mercaptoethanol, 500 mM ammonium chloride, 100 mM magnesium acetate) and lysed at 4 $^{o}$C using glass beads and a Bead Beater (BioSpec, Inc) for 10 min as previously described (Browne et al. 2013). The whole cell extracts

(WCE) were clarified by centrifugation at 20,000g for 15 min at 4 oC and a 200 µL aliquot of the cleared WCE was stored at -80 $^{o}$C. The remaining cleared WCEs were overlaid onto a 5/20% discontinuous sucrose gradient prepared in wash buffer. The gradients were centrifuged at 28,000 RPM using a SW-41 swinging bucket rotor for 18 h at 4 $^{o}$C. The supernatant was discarded and the ribosome pellet was resuspended in ice cold 1 mL standard buffer (10 mM Tris pH 8, 5 mM beta-mercaptoethanol, 50 mM ammonium chloride, 5 mM magnesium acetate) and centrifuged for 10 min at 10,000g at 4 $^{o}$C. The pellet was discarded and the ribosome suspension was stored at -80 $^{o}$C. For cryoEM analysis, ribosomes were purified as above after shifting the cells grown in glucose to glycerol and taking aliquots at the following time points: 0Min, 30Min, 60Min, 120Min, 240Min, and 450Min.

**iTRAQ labeling**: The total protein concentration of all ribosome suspensions were determined using Bradford assay according to the manufacturer protocol (Sigma Aldrich, St. Louis, MO. Catalog # B6916-500ML). Fifty micrograms of total protein from each growth condition was mixed with 50 ng of bovine serum albumin (Thermo Scientific, #23209) as an internal standard. Each protein sample was acetone precipitated and resolublized in 25 µL iTRAQ dissolution buffer (500 mM triethylammonium bicarbonate, 0.1 % sodium dodecyl sulfate). The proteins were reduced with tris(2-carboxyethyl)phosphine at 60 $^{o}$C for 60 min and the cysteines were derivatized with methyl methanethiosulfonate at RT for 10 min. All samples were digested with sequencing grade modified trypsin (1:50; Promega Corporation, Catalog # V5111) overnight at 37 $^{o}$C. An equal fraction of the tryptic digest of ribosomes from the 3 replicates grown in YPD at 30 $^{o}$C were pooled separately and used as a control for the

iTRAQ experiments. Ten µg from each replicate tryptic digested sample and pooled control were used for iTRAQ labeling. The iTRAQ labeling reagents were resolublized in 150 µL anhydrous ethanol (Sigma Aldrich, St. Louis, MO. Catalog # E7023-500ML). Seventeen µL of iTRAQ reagents were added to each 10 µg sample and the pooled control, incubated with shaking for 1 h at room temperature on Eppendorf Thermomixer R, pooled, frozen, lyophilized, resolublized in X ul of buffer A (0.1 % formic acid in HPLC-grade water), and stored at -80 oC.

**Liquid chromatography and mass spectrometry**: The iTRAQ labeled samples were analyzed with MudPIT as previously described (Browne et al. 2013). Precursor ions were analyzed in the Orbitrap mass analyzer followed by 4 CID fragment ion scans in the ion trap and 4 HCD fragment ion scans (normalized collision energy = 45%) in the Orbitrap.

**iTRAQ data analysis**: The data analysis workflow essentially mirrored the workflow described in Chapter II. Briefly, RAW files generated by the MudPIT experiments were searched using the Sequest database search engine running under Proteome Discoverer v1.4 (Thermo Scientific) against a forward and reverse yeast protein database (S.cererevisiae_orf_trans_all_SGD.fasta.6718) with appended common contaminant sequences (Eng et al. 2008; Eng, McCormack, and Yates 1994). Protein assembly and reporter ion quantitation and statistical analysis were done using ProteoIQ (Premier BioSoft Inc). Principal component analysis was done using princomp function in R and the PCA plot was generated using Scatterplot3d package (R Core Team 2015; Ligges and Mächler 2003). Boxplots were generated in RStudio. All python and R scripts used in this study will be made available on request.

**Multiple Reaction monitoring**: Proteotypic peptides were selected for targeted quantitation from a database of identified peptides in the MudPIT experiments. Transitions for unscheduled scout experiments were selected based upon NIST and GPM spectral libraries. Fifty µg of the purified ribosomes were digested with sequencing grade modified trypsin (1:50; Promega Corporation, Catalog # V5111) and desalted essentially as described (Browne et al. 2013). Peptides were eluted using an elution buffer composition of 50% Acetonitrile, 0.1% Trifluoroacetic acid. Peptides were analyzed using a 90 min scheduled SRM analysis. Briefly, peptides were autosampled onto a 200 mm by 0.1 mm (Jupiter 3 micron, 300A), analytical column coupled directly to an TSQ-Vantage (ThermoFisher) using a nanoelectrospray source and resolved using an aqueous to organic gradient (1-45% Buffer B) at X ul/min flow rate. Using series of unscheduled scout runs to determine retention times and transitions to monitor, a scheduled instrument method encompassing a 10 min window around each retention time along with calculated collision energies was created using Skyline (MacLean et al. 2010).Q1 peak width resolution was set to 0.7, collision gas pressure was 1 mTorr, and utilized an EZmethod cycle time of 5 s. The resulting RAW instrument files were imported into Skyline for peak-picking and quantitation (MacLean et al. 2010). The peak areas of the transitions were exported and further analysis done in Microsoft Excel. Sum of the peak areas of all the transitions of a given peptide, peptide peak area, was used as the quantitative measure of abundance for the peptide. The average of peptide peaks areas of all the peptides from a given protein, protein peak area, was used as the quantitative measure of abundance of the protein. The average protein peak areas of single copy ribosomal protein RPL5 was used as control. For differential

analysis, in the first step a ratio of peak area of the test protein to the peak area of the control was calculated across all samples. In the next step, two sample *t*-test with alpha level 0.05 was performed with the ratios to test for statistical significance. Finally, fold change was calculated by ratioing the average of ratios. The calculated was fold changes were log2 transformed.

**Cryoelectron Microscopy (in collaboration with the Joachim Frank lab)**: For each specimen of the time series, ribosome samples were applied to Holey carbon-coated Quantifoil copper grids, freeze-plunged using the Vitrobot Mark IV freeze-plunger (FEI, Portland, Oregon), and then visualized in an FEI Tecnai F20 electron microscope at 200 kV acceleration voltage and 5,000x magnification, using a 4k x 4k CCD camera (Gatan, Pleasanton, CA) and automated data collection employing the programs Leginon and Appion (Grassucci, Taylor, and Frank 2008; Suloway et al. 2005). Each pixel corresponds to 2.25A on the object scale. A total number of 260,440 particles were selected from 2,661 micrographs. Of these, 159,654 were verified using a work-flow written in Arachnid, and processed using the program RELION, which combines maximum likelihood-based classification with reconstruction, as well as a novel convergence analysis that finalizes the classification results (Bo Chen, Shen, and Frank 2014; Scheres 2012). For a comprehensive analysis of the time series, all data were pooled together so that increase and decrease of each sub-population could be effectively studied, and the maximum number of particles was available for the 3D reconstruction of each class.

**GOzilla**: GOzilla is a custom *Python* script. It uses GO Slim database from Saccharomyces cerevisiae Genome Database (SGD) (Cherry et al. 1998; Christie et al.

2004). In the first step COMPzilla creates a dictionary with GO terms as the keys and the proteins that have the given GO term associated with them in the GO Slim database as the values. In the next step, GOzilla creates a dictionary in which GO terms are still the keys, but values are fold changes corresponding to the proteins that were mapped to GO terms in the first step. In the third step, GOzilla creates a list all the fold changes in the experiment that will be considered the population of fold changes for statistical testing. Finally, GOzilla compares the fold change distribution associated with the GO terms with population fold change distribution using two sample *t*-test of independence and two sample Kolomogorv-Smirnov test. GOzilla exports the results of the two tests in separate tab delimited text files, in which first column contains the GO terms, the second column *t*-statistics or *ks*-statistics, and the third column contains the corresponding *p*-value (Source code in Appendix R).

**COMPzilla**: COMPzilla is a custom *Python* script. It uses CYC2008 2.0, a manually curated database of biomolecular complexes in yeast to identify complexes that are differentially present (Pu et al. 2007; Pu et al. 2009). In the first step COMPzilla creates a dictionary with complex names as keys and the proteins that constitute the complex as values. In the next step, COMPzilla creates a dictionary in which complex names are still the keys, but values are mapped fold changes of the proteins that constitute the complex. In the third step, COMPzilla creates a list all the fold changes in the experiment that will be considered the population of fold changes for statistical testing. Finally, COMPzilla compares the fold change distributions associated with protein complexes with population fold change distribution using two sample *t*-test of independence and two sample Kolomogorv-Smirnov test. COMPzilla exports the results

of the two tests in separate tab delimited text files, in which first column contains the complex names, the second column *t*-statistics or *ks*-statistics, and the third column contains the corresponding *p*-value. (Source code in Appendix S)

## Results

### Ribosomal proteins abundances are regulated in response to environmental stimuli but the abundances of all RPs do not change to the same extent.

To investigate the regulation of abundances of ribosomal proteins in response to environmental stress, we reanalyzed our previously published dataset using *Python* scripts *GoZilla.py* and *CompZilla.py* ((Source codes in Appendices R and S). The whole cell extract analyzed in this study were the sources for the purified ribosomes for this study (Samir et al. 2015). *GoZilla* identifies the Gene Ontology terms that are either downregulated or upregulated in a gene expression data. It uses the Go Slim database downloaded from Saccharomyces genome database for looking up GO terms (Cherry et al. 1998; Christie et al. 2004). *CompZilla* identifies the differentially regulated biomolecular complexes in gene expression data. It uses a manually curated database of yeast biomolecular complexes to lookup their constituents (Pu et al. 2007; Pu et al. 2009).

In the three stimuli used in the previous study, most of the proteins with structural constituent of ribosome GO terms were downregulated (Fig. 12A). Similarly most of the components of 60S and 40S ribosomal subunits were downregulated too (Fig. 12B). However, the $\log_2$ transformed fold changes were not consistent for all of the ribosomal proteins. This suggested that at least some of the ribosomal proteins were being

differentially regulated compared to others. This is in agreement with the predictions of both ribosome filter hypothesis and the depot hypothesis (Mauro and Edelman 2002; Ray, Arif, and Fox 2007; Mazumder et al. 2003). However, this does not exclude the possibility that the changes in ribosomal protein abundances were independent of the changes in the protein composition of ribosomes themselves.

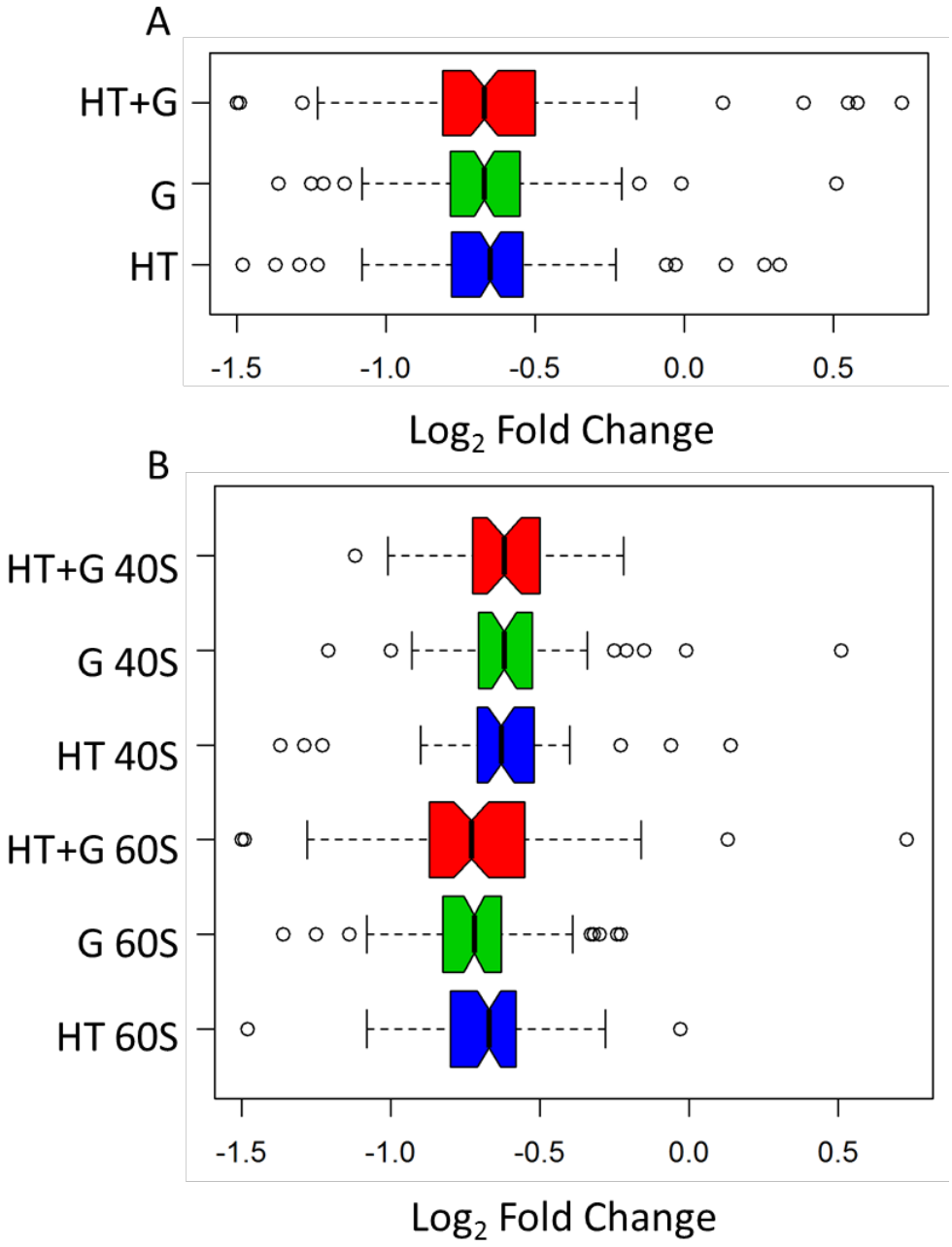



**Figure 12 -** : Analysis of whole cell extract quantitative protoemics data using GoZilla and CompZilla.

A) A boxplot showing the fold change distribution of proteins with structural constituent of ribosome GO term. GoZilla was used to generate the fold change distribution. Boxplots were generated in RStudio. B) Fold change distributions of 40S and 60S subunit proteins. CompZilla was used to generate the fold change distribution. Boxplots were generated in RStudio.

**Quantitative proteomics analysis of protein abundances in purified ribosomes.**

To directly measure the changes in the protein composition of ribosomes, we focused on two environmental stimuli – (1) growth in rich media with glucose as carbon source at 30 $^{o}$C, and (2) growth in rich media with glycerol as carbon source at 30 $^{o}$C. We purified the ribosomes using a discontinuous sucrose gradient centrifugation. We analyzed the samples using iTRAQ labeling followed by liquid chromatography tandem mass spectrometry. We identified 135 ribosomal proteins, 131 of which were quantitated in the three replicates. Since the ribosomes were purified from the same whole cell extracts used in the previous study, we compared the fold changes of ribosomal proteins in purified ribosomes to that in the whole cell extracts (Samir et al. 2015). Surprisingly, a correlation matrix analysis revealed that there was no correlation between the two fold changes (Fig. 13A-D). There was good correlation in the ribosomal protein levels between the replicates, both in whole cell extracts and purified ribosomes (Fig. 14A-F). This suggested that the lack of correlation was not due to a noisy data with high variance.

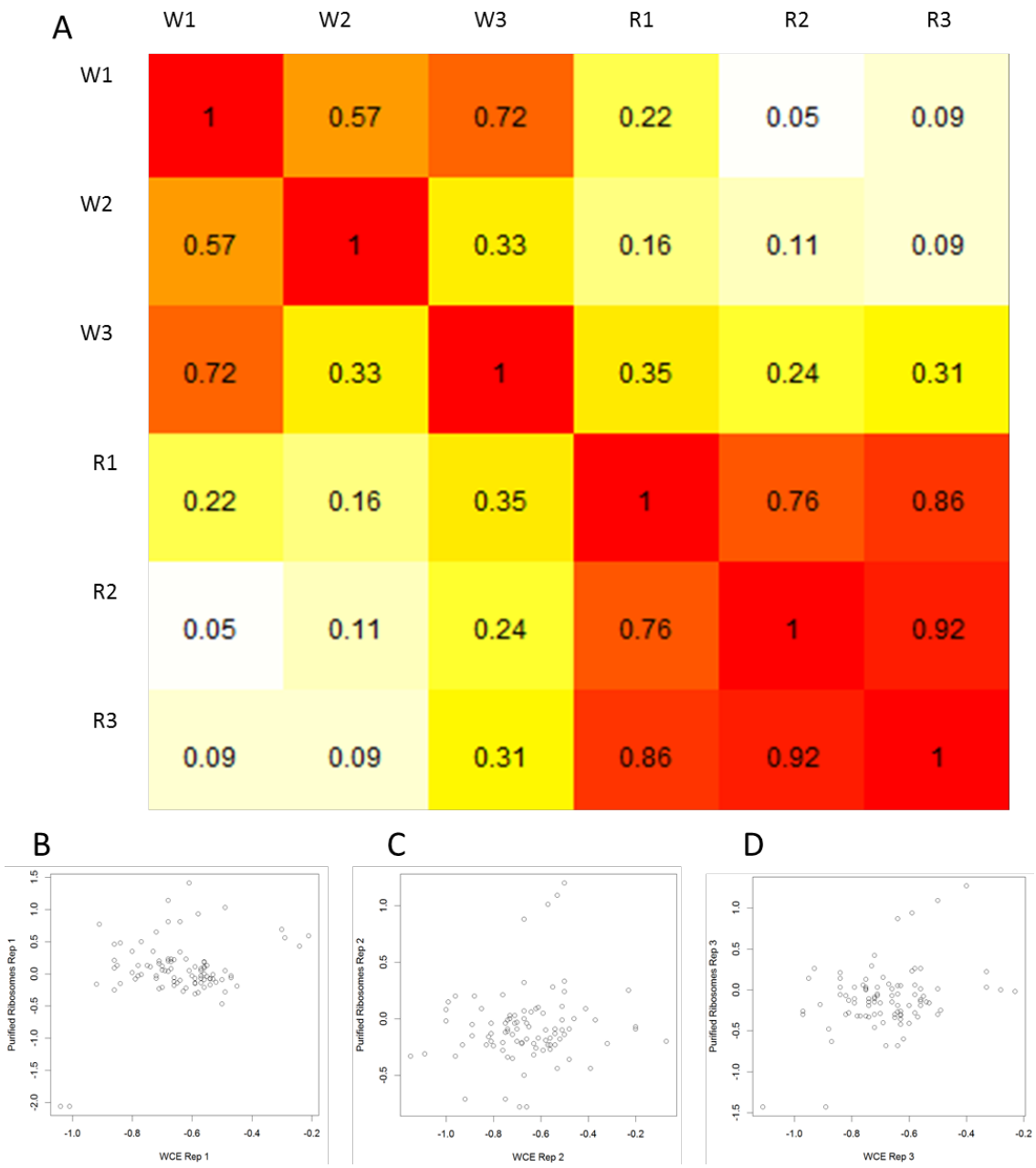



**Figure 13** – Correlation analysis between the changes in whole cell extracts and purified ribosomes

There are three replicates each corresponding to WCE and purified ribosomes. A) Cross0correlation matrix, numbers represent Pearson's R, W# represent whole extracts, R# represent purified ribosomes. B-D) Scatter plots showing relationship between fold changes in whole cell extract (X-axis) and purified ribosomes (Y-axis) Three replicates are depicted.

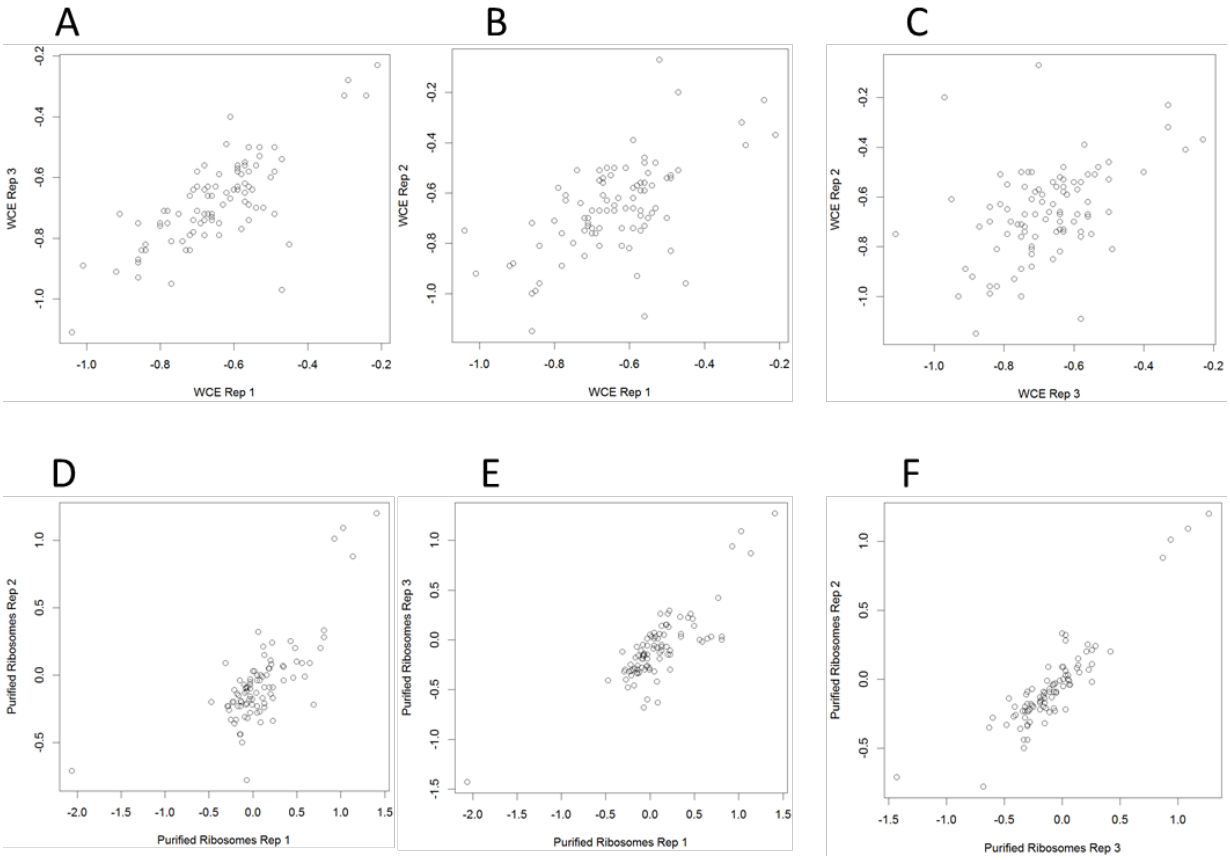



**Figure 14 -** Scatterplots showing reproducibility among replicates of whole cell extracts and purified ribosomes

Sctterplots were generated in RStudio. A-C) Scatterplots of whole cell extracts. Replicate 1 vs Replicate 2, Replicate 1 vs Replicate 3, and Replicate 3 vs Replicate 2. D-F) Scatterplots of purified ribosomes. Replicate 1 vs Replicate 2, Replicate 1 vs Replicate 3, and Replicate 3 vs Replicate 2.

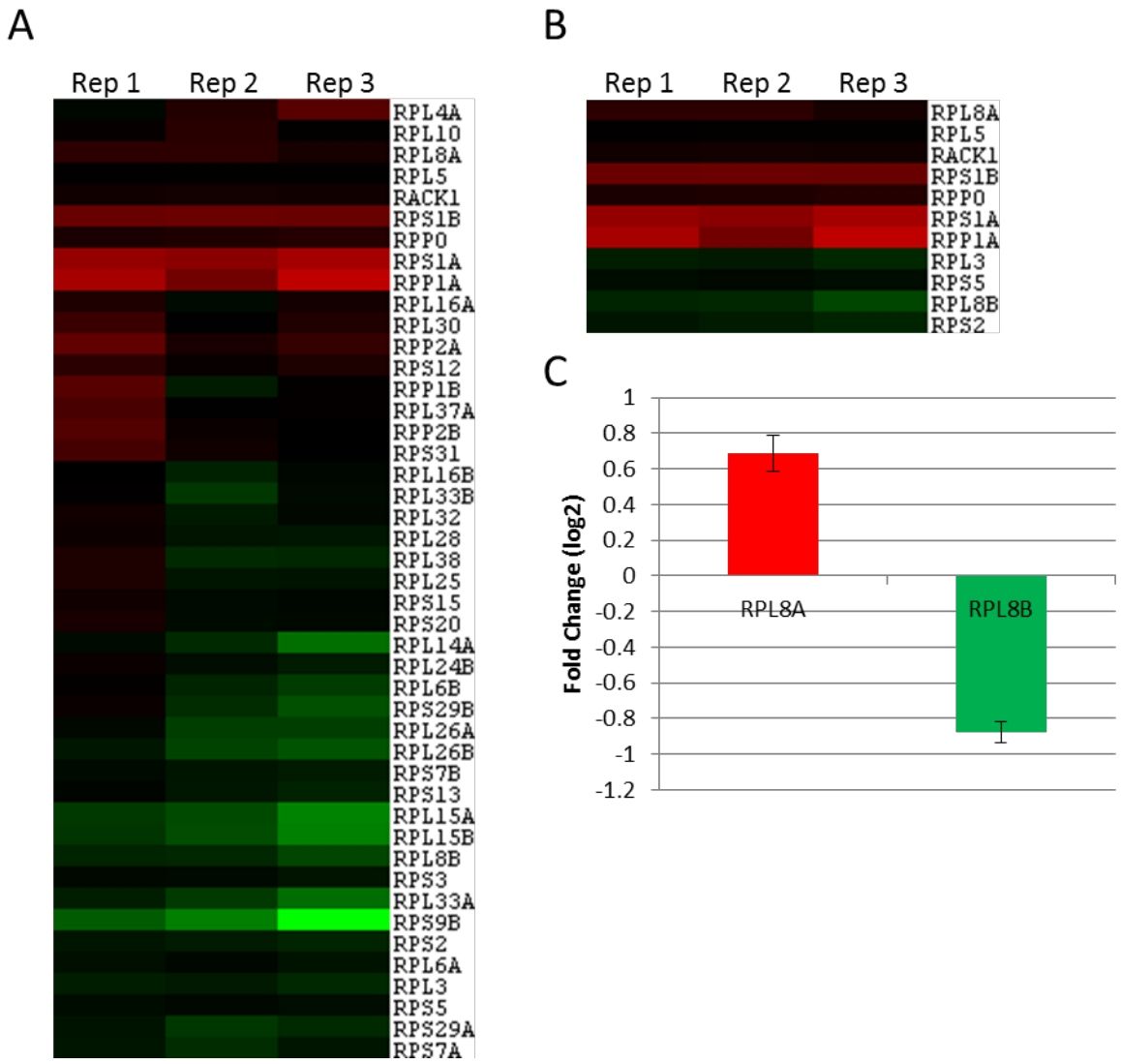



**Figure 15** – Quantitation of ribosomal proteins in purified ribosomes using quantitative mass spectrometry

A) 45 ribosomal proteins quantitated using at least one unique peptides in iTRAQ experiments. B) 11 Differentially present ribosomal proteins in purified ribosomes identified using *t*-test *p-value* less than 0.05. C) The amounts of RPL8A and RPL8B were validated using multiple reaction monitoring approach in three independently purified ribosomes. These were different samples from the ones used in iTRAQ.

To identify differentially present ribosomal proteins, we used *t*-test of independence with alpha level of 0.05. Since there are minimal sequence differences between the paralogs, to reliably quantify the paralog specific changes we reanalyzed

the mass spectrometry data to use only the unique peptides for quantitation. We quantitated 45 ribosomal proteins with at least one unique peptide, 11 of which were differentially present in the purified ribosomes (Fig. 15A-B). This included a paralog pair, Rpl8a and Rpl8b (Fig. 15C). We validated the iTRAQ data using multiple reaction monitoring.

**Using CryoEM to detect changes in the ribosomal protein composition over time.**

Quantitative proteomics is a population based technique that could not address superstoichiometric and substoichiometric models for changes in the protein composition of ribosomes. To determine the changes in the protein composition of ribosomes consistent with superstoichiometric and substoichiometric models, we used cryoEM in collaboration with the Joachim Frank lab at Columbia University. We focused on the changes in 80S ribosomes. We identified 3 populations of ribosomes, (1) a population of complete 80S ribosomes that has the full complement of ribosomal proteins, (2) a population of 80S ribosomes missing Rpl10 (ΔuL16), and (3) a population of ribosomes missing both Rpl10 and Rps1 (ΔuL16 ΔeS1) electron densities (Fig. 16A). We did not observe ribosomes with superstoichiometric composition of ribosomal protein(s) in our data.

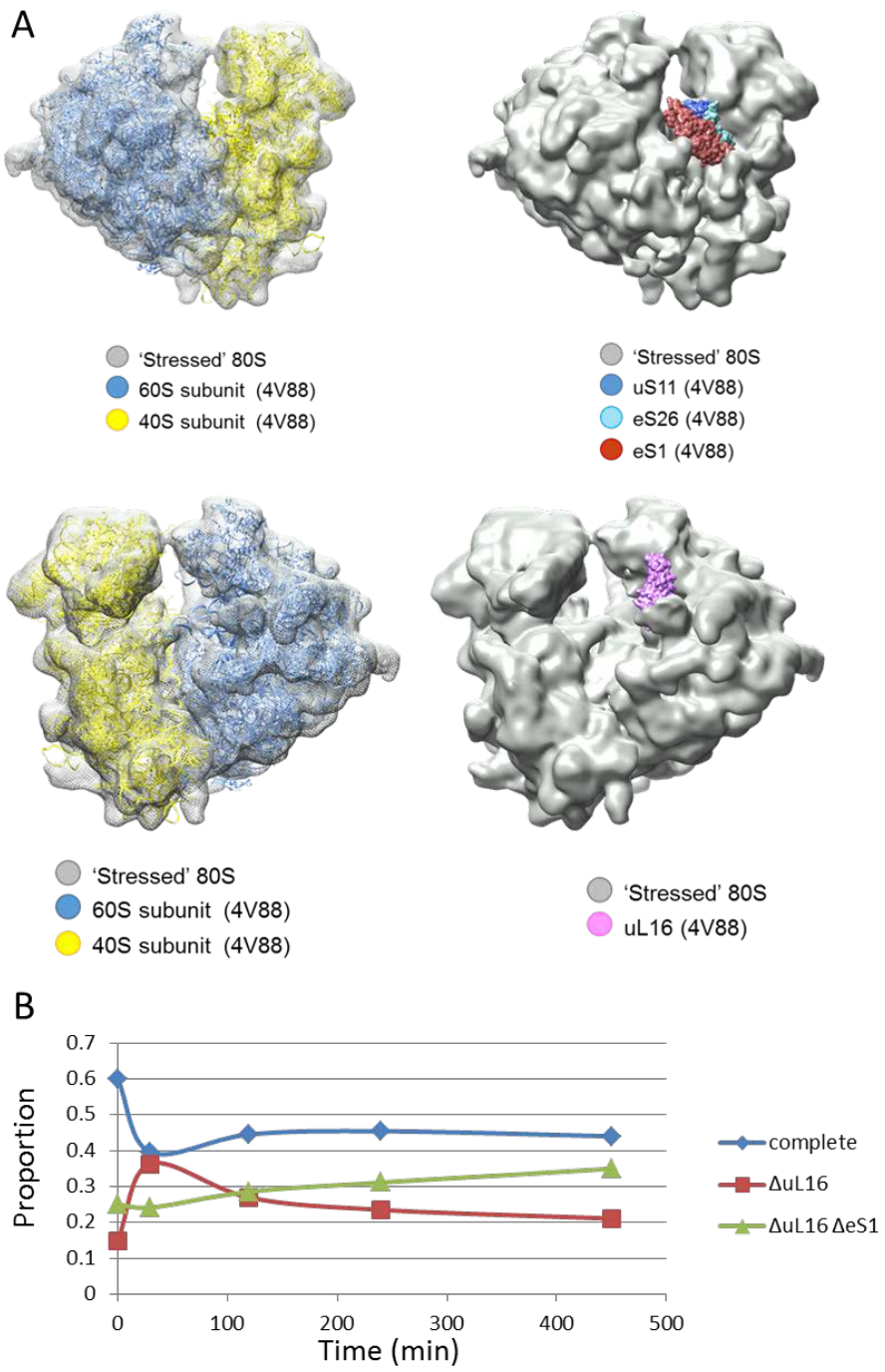



**Figure 16 -** Cryo-EM analysis

A) Electron densities corresponding to Rpl10 (uL16) and Rps1 (eS1) in cryo-EM structures. Substoichiometric ribosomes lacking these proteins were identified. B) Time-course analysis to measure the dynamic changes in the proportions of ribosomes missing uL16 or both uL16 and eS1.

To study the kinetics of change in the composition of 80S ribosomes, we performed a time-course experiment to track the proportions of these sub-populations. The cells were grown in with glucose as the carbon source. At time 0 minutes, the cells were spun down and resuspended in media containing glycerol as the carbon source. We aliquoted cells at 0, 30, 60, 120, 240, and 450 minutes after shifting from glucose to glycerol. We purified ribosomes from these cells and used cryo-EM to determine the relative proportions of the three structures (Fig. 16B). The proportion of complete ribosomes decreased sharply within the first 30 minutes of the shift. There was a similar increase in the proportion of ΔuL16 ΔeS1 ribosomes in the same time frame. After 60 minutes, the proportion of complete ribosomes started to recover with a concomitant decrease in the ΔuL16 ΔeS1 ribosomes. However, the proportions of complete and ΔuL16 ΔeS1 ribosomes never recovered to the initial level in the time frame used in this study. The proportion of ΔuL16 ribosomes continued to steadily increase throughout our experiment. However, the rate of increase was minimal. The substoichiometric composition is consistent with both the depot hypothesis and the ribosome filter hypothesis.

**Paralog specific roles of Rpl8a and Rpl8b in translation using null mutants.**

We used polysome profile analysis to study the paralog specific roles of Rpl8a and Rpl8b in global translational control. We used the *rpl8a* and *rpl8b null* mutants that had been previously described (Winzeler et al. 1999). We used the diploid wild type (BY4743) and *null* mutant strains (Baker Brachmann et al. 1998; Winzeler et al. 1999). Similar to the proteomics analysis, we grew cells with either glucose or glycerol as carbon source (Fig. 17 and 18). Polysome profiles of *rpl8a* cells showed a large

increase in 40S peak and shoulders on the 80S and polysome peaks (Fig. 17B). The shoulder defect was rescued by either adding back Rpl8a or overexpressing Rpl8b from their native promoters (Fig. 17B, 17D, 17E). However, the 40S peak defect was rescued only by adding back Rpl8a (Fig. 17B, 17D, 17E). Polysome profiling of *rpl8b cells* did not show a difference from wild type (Fig. 17C). When the cells were grown with glycerol, *rpl8a* cells showed a very prominent 40S peak, which was rescued by adding back Rpl8a (Fig. 18B, 18D). This defect was not rescued by overexpression of Rpl8b (Fig. 18E). In glycerol, *rpl8b* cells too showed a larger 40S peak that was rescued by Rpl8b but not by Rpl8A (Fig. 18C, 18F, 18G).

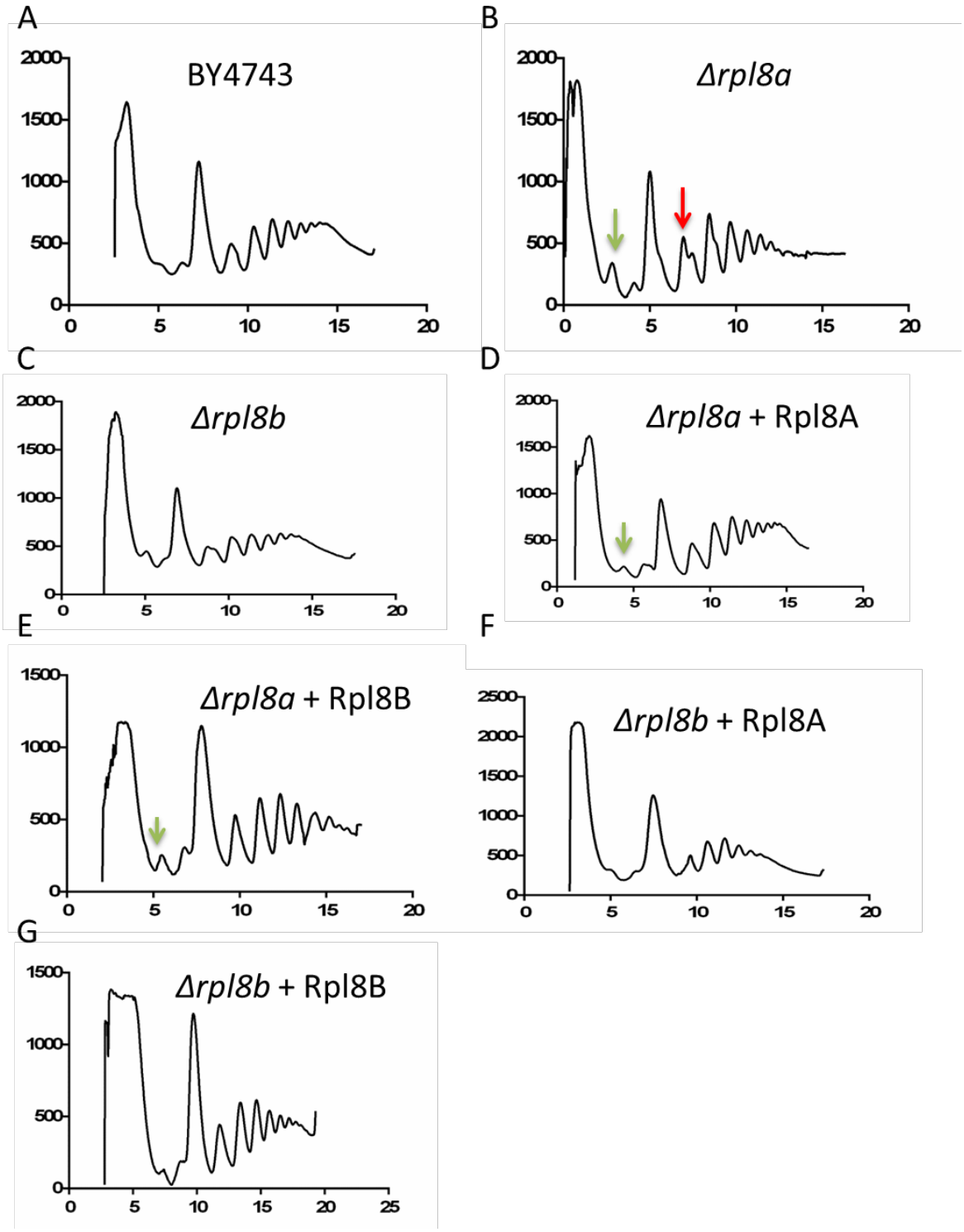



**Figure 17** – Polysome profiles with glucose as carbon source

A) Polysome profile of WT diploid strain *BY4743*. B) Polysome profile of *rpl8a diploid null* mutant. C) Polysome profile of *rpl8b diploid null* mutant. D) Polysome profile of *rpl8a diploid null* mutant with Rpl8a on a plasmid expressing from native promoter. E) Polysome profile of *rpl8a diploid null* mutant with Rpl8b on plasmid expressing from native promoter. F) Polysome profile of *rpl8b diploid null* mutant with Rpl8a on a plasmid expressing from native promoter. E) Polysome profile of *rpl8b diploid null* mutant with Rpl8b on plasmid expressing from native promoter.

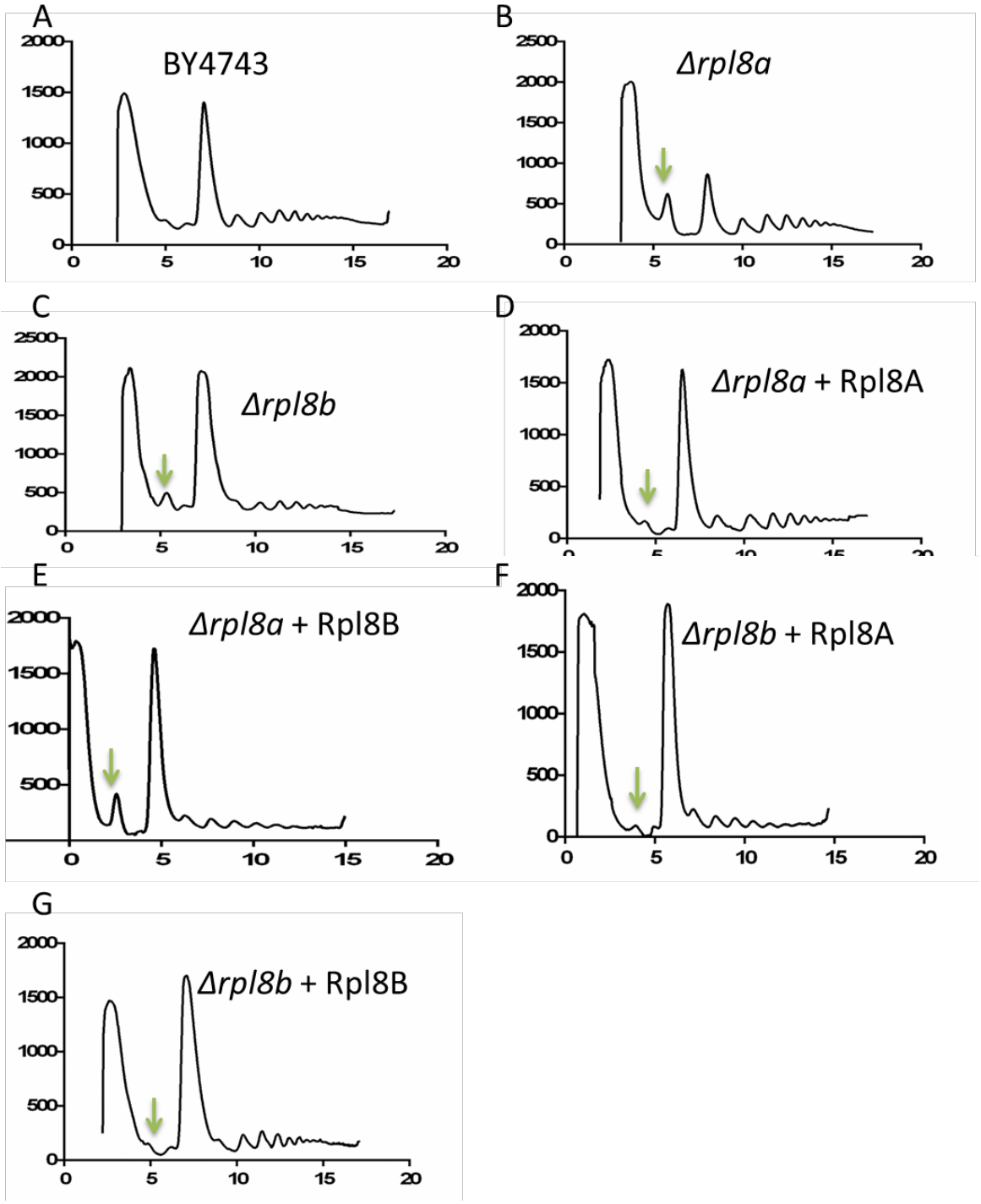



**Figure 18** – Polysome profiles with glycerol as carbon source

A) Polysome profile of WT diploid strain *BY4743*. B) Polysome profile of *rpl8a diploid null* mutant. C) Polysome profile of *rpl8b diploid null* mutant. D) Polysome profile of *rpl8a diploid null* mutant with Rpl8a

on a plasmid expressing from native promoter. E) Polysome profile of *rpl8a diploid null* mutant with Rpl8b on plasmid expressing from native promoter. F) Polysome profile of *rpl8b diploid null* mutant with Rpl8a on a plasmid expressing from native promoter. E) Polysome profile of *rpl8b diploid null* mutant with Rpl8b on plasmid expressing from native promoter.

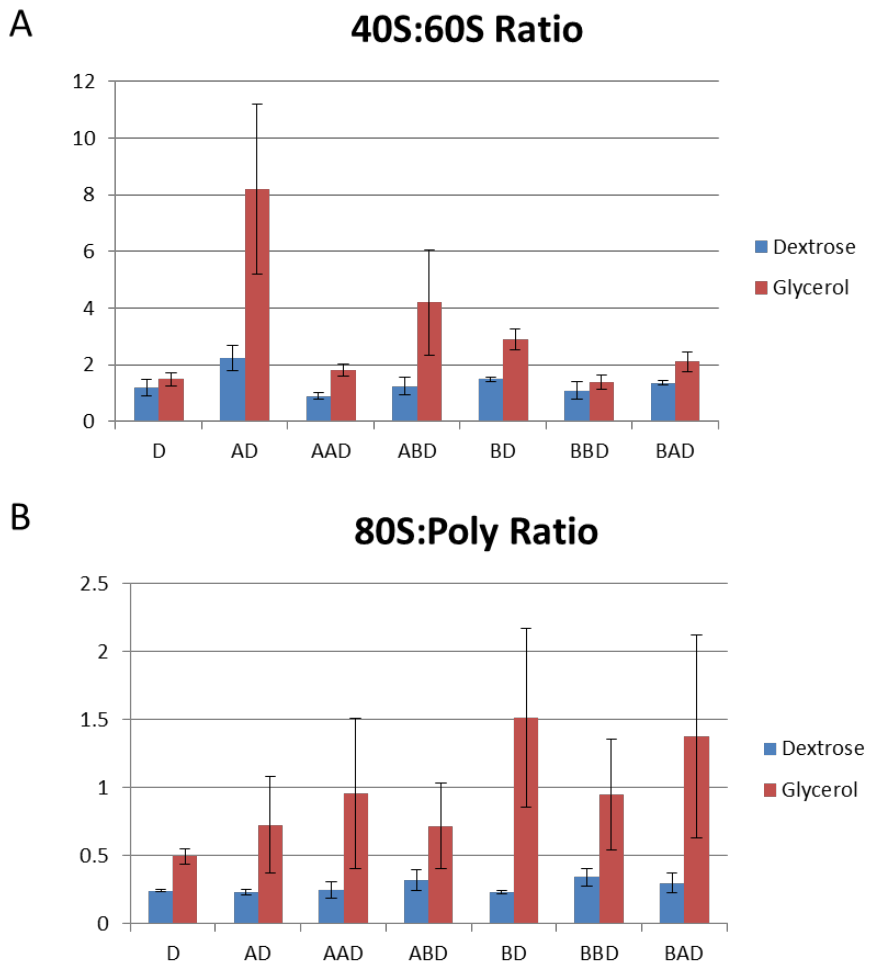



**Figure 19** – Quantification of peak areas ratios of polysome profiles

A) Ratio of 40S:60S peak areas. B) Ratio of 80S:Polysome peak areas. D is *BY4743*, AD is *rpl8a diploid null*, AAD is *rpl8a diploid null* with Rpl8a, ABD is *rpl8a diploid null* with Rpl8b on a plasmid, BD is *rpl8b diploid null*, BBD is *rpl8b diploid null* with Rpl8b, BAD is *rpl8b diploid null* with Rpl8a on a plasmid.

We further analyzed the polysome profiles quantitatively by measuring the peak areas of the 40S, 60S, 80S, and polysome peaks. We used two parameters – (1) the ratio of 40S:60S peak areas (40/60 ratio), and (2) the ratio of 80S:Polysome peak areas (80/Poly ratio) (Fig. 19). Both *rpl8a* and *rpl8b* cells showed elevated 40/60 ratios with

either glucose and glycerol as carbon sources (Fig. 19A). In *rpl8a* cells. the ratios were rescued to the wild type levels by adding back Rpl8a (Fig. 19A). Although the overexpression of Rpl8b decreased the ratios, but it did not reach the wild type levels (Fig. 19A). In *rpl8b* cells too showed a similar pattern, with adding back Rpl8b rescuing the defect (Fig. 19A). Overexpression of Rpl8a was only able to partially rescue the defect (Fig. 19A).

Analysis of 80S/Poly ratios showed no defectin either *rpl8a* or *rpl8b* cells when grown in glucose (Fig. 19B). Overexpression of Rpl8a or Rpl8b in the mutant cells did not lead to a defect either (Fig. 19B). When cells were grown in glycerol, there was an elevation in the ratio for *rpl8b* cells but not for the *rpl8a* cells (Fig. 19B). The elevated 80S/Poly ratio defect was rescued by adding back Rpl8b but not by overexpression of Rpl8a (Fig. 19B). Taken together, the polysome profile analysis suggested paralog specific roles for Rpl8a and Rpl8b proteins. This is consistent with the ribosome filter hypothesis.

**Discussion**

Presence of ribosomes with substoichiometric composition of ribosomal proteins presents intriguing possibilities. In *E. coli*, ribosomes lacking a 43 nucleotide segment on the 3’ end of the 16S rRNA selectively translates leaderless mRNA. This would be an example of a ribosome with substoichiometric composition, in this case a rRNA, acting as a filter for translating a specific class of mRNA. Do the ribosomes missing Rpl10, Rps1, or both translate a special class of mRNAs? It is also possible that Rpl10 and Rps1 have extra-ribosomal functions similar to RPL13A in GAIT complex. There dissociation from the ribosome to perform their extra-ribosomal function may leave

behind ribosomes lacking their electron densities. Another intriguing possibility is that the dissociation of Rpl10 and/or Rps1 acts as a break. This is consistent with the observation that the proportion of ribosomes missing either of the two proteins increases rapidly immediately after shifting the carbon source (Fig. 16B). Since cells have to adapt to a new environment, they may pause for a time before making new proteins. Furthermore, the proportion of complete ribosomes never recovers to the pre shift stage during the time frame of this study. Since yeast cells grow very slowly with glycerol as carbon source compared to glucose, the ribosomes lacking one or both of the ribosomal proteins may be part of a non-translating reserve pool. Although we cannot differentiate between the three possibilities, the very presence of substoichiometric ribosomes suggests alternative biological models that would need to be addressed in future studies.

Although the change in composition of ribosomes with the changes in the growth condition was not completely unexpected, we provide experimental evidence that it does occur, at least in *S. cerevisiae*. The paralog specific roles of Rpl8a and Rpl8b observed in the polysome profiles provide direct evidence in favor of the ribosome filter hypothesis. In planned follow up experiments, we are using ribosome footprint profiling with the wild type and mutant cells to identify the mRNAs whose translation is differentially affected by either of the paralogs (Ingolia 2014). It would be very interesting to see if there are sequence features on the mRNAs under paralog specific translational control. A recent study in mouse has found RNA secondary structures in 5' UTRs of a subset of *Hox* mRNAs (Xue et al. 2015). These RNA structures resemble internal ribosome entry sites required for the cap-independent translation initiation

mechanism. They are expected to recruit ribosomes to the mRNA through a mechanism dependent upon Rpl38 (Xue et al. 2015).

A number of single copy ribosomal proteins are differentially present in the purified ribosome samples in the proteomics experiment but not in the cryo-EM experiments focused on 80S ribosomes. There could be three possibilities that can explain these observations – (1) confounding of the mass spectrometry quantitation by impurities, unassociated 40S and 60S subunits, biogenesis or degradation intermediates, in the sample prep, (2) cofounding of the mass spectrometry quantitation by unidentified post-translational modifications, and (3) the changing composition in free 40S and 60S subunits that were excluded from cryo-EM analysis.

In our experiments, we pelleted down all of the ribosomes including 40S, 60S, and 80S. Our protocol would also have pelleted the ribosomes in the intermediate stages of biogenesis or degradation. We cannot differentiate between the tryptic peptides coming from the different sources in the sample. Each of these peptides contributes to quantification and potentially confounds our results. In cryo-EM, we focused only on the 80S ribosomes. We were able to filter out all the other sources of variation from our analysis. Therefore, cryo-EM provides a cleaner data for the single copy ribosomal proteins. However, cryo-EM cannot differentiate between the different paralogs at the resolutions routinely achieved through this approach. This makes quantitative proteomics and cryo-EM complementary techniques to address these different questions.

A second reason for the changes detected in the single copy ribosomal proteins could be the confounding by the unidentified post-translational modifications. Ribosomal proteins are known to be post-translationally modified (Ohn et al. 2008; Spence et al. 2000; W. M. Anderson, Grundholm, and Sells 1975; Kruiswijk et al. 1978; Kaerlein and Horak 1976; Arragain et al. 2010; Nesterchuk, Sergiev, and Dontsova 2011; Xirodimas et al. 2008; Arnold et al. 1999; Thomas et al. 1982). The post-translational modification of peptides confounds mass spectrometry based quantitation if the modified peptides have been not been identified in the experiment. This is because the modification changes the mass as well as the retention time for a chromatography run. This leads to absence of signal that can be misinterpreted as differential presence. To minimize this possibility, use of two or more peptides for quantitation is recommended because the probability of both peptides being modified simultaneously is considered lower than an individual peptide. Although these precautions can minimize the chances of errors in quantitation due to post-translational modifications, it cannot completely rule them out.

A third reason for the changes in the single copy ribosomal protein could be that the proportions of ribosomes missing these ribosomal proteins are changing between the two growth conditions. Although we did not observe the 80S structures missing these ribosomal proteins in cryo-EM analysis, they might be present in the ribosomal subunits or intermediates of biogenesis or degradation.

In conclusion, we have showed the changing protein compositions of ribosomes using two complementary approaches. Paralog specific changes are consistent with the ribosome filter hypothesis. The changes in the single copy ribosomal proteins observed in the quantitative proteomics study and cryo-EM are consistent with both the depot

hypothesis and the ribosome filter hypothesis. The planned future studies are expected to shed light on the functional significance of the changing compositions.

# Chapter IV

## Quantitative Proteomics Analysis of Human Myotonic Dystrophy Skeletal Muscles Reveals Specific and Common Modules of Differentially Expressed Proteins.


### Abstract

Myotonic dystrophy, a form of muscular dsystrophy, is an autosomal dominant multi-systemic disorder caused by the expansion of nucleotide repeats. There are two types of myotonic dystrophy. Myotonic dystrophy type 1 (DM1) is cause by a CTG trinucleotide repeat expansion in the 3' untranslated region of dystrophia myotonica-protein kinase (DMPK) gene. Myotonic dystrophy type 2 (DM2) is caused by a CCTG tetranucleotide repeat expansion in the first intron of zinc finger 9 (ZNF9) gene. The expression of the repeat expansions in both cases lead to the nuclear accumulation of RNA granules, which sequester RNA processing factors. This RNA toxicity is thought to be the cause of the disease symptoms. However, the effect of the repeat expansions on the proteome is poorly understood.

To address this, we quantified the proteomes of the skeletal muscles of myotonic dystrophy patients and healthy volunteers to identify differentially regulated proteins. We used iTRAQ labeling followed by liquid chromatography tandem mass spectrometry for protein quantitation. Skeletal muscles from 5 healthy volunteers, 7 DM1 patients and 6 DM2 patients were used in this study. We quantitated 3575 proteins across all the samples. We used one way ANOVA, with the Benjamini Hochberg procedure for controlling false discovery rate in multiple comparisons, to identify differentially

regulated proteins. We identified 30 proteins upregulated and 4 proteins downregulated in both DM1 and DM2. We found 154 proteins to be upregulated and 218 proteins to downregulated uniquely in DM1 patients. Pathway analysis of these proteins revealed biochemical pathways that appear to be affected by the repeat expansions.

## Introduction

Myotonic dystrophy (DM) is an autosomal dominant multi-systemic disorder caused by the expansion of CTG or CCTG repeat elements in DMPK or ZNF9 genes, respectively. DM caused by the expansion CTG repeat in the 3'untranslated region of DMPK gene is called myotonic dystrophy type 1 (DM1) (Brook et al. 1992; Y. H. Fu et al. 1992; Mahadevan et al. 1992). It is also called Steinert's disease and congenital myotonic dystrophy (Machuca-Tzili, Brook, and Hilton-Jones 2005). It is the more severe form of DM. DM caused by the expansion of CCTG repeat in the first intron of ZNF9 gene is called myotonic dystrophy type 2 (DM2) (Ranum et al. 1998; Liquori et al. 2001). It is also called the proximal myotonic dystrophy (PROMM). This is a relatively mild form of DM (Machuca-Tzili, Brook, and Hilton-Jones 2005). DM1 and DM2 are thought to be caused by accumulation of toxic RNA (J. E. Lee and Cooper 2009; Osborne and Thornton 2006; Thornton 2014). DM symptoms include myotonia, cataracts, neurological disorders and heart conduction defects.

DM1 was first described more than hundred years ago (Machuca-Tzili, Brook, and Hilton-Jones 2005). It was the third example of a disease caused by repeat expansions in 1992 (Thornton 2014; Brook et al. 1992; Mahadevan et al. 1992; Y. H. Fu et al. 1992). The number of CTG repeat in general human population is variable. The number of repeats in healthy individuals lies between 5 and 37 (Thornton 2014). In DM1

patients the number of repeats exceeds 50 and can even be more than 3000 (Thornton 2014). The number of repeats correlates with the degree of severity of the disease. It also negatively correlates with the age of onset, the larger of number of repeats being associated with an earlier age of onset (Redman JB et al. 1993; Temmerman et al. 2004). DM1 can present in either congenital or adult onset form. This depends on the number of repeats in an individual.

DM2 was first described in 1994 as the myotonic dystrophy that lacked the CTG repeat expansion described two years earlier (Thornton, Griggs, and Moxley 1994). A number of DM patients were soon found to lack the CTG expansion (Ricker K et al. 1995; Ricker et al. 1994; Meola et al. 1996; Udd et al. 1997). This form of the disease was initially called proximal myotonic myopathy because of the involvement of proximal muscles, in contrast to the distal muscles in the previously described form of DM (Udd et al. 1997; Moxley III 1996; Ricker et al. 1994). It was later renamed as DM2 to signify the form of DM that lacked CTG repeat expansion. DM2 is an adult onset disease. A congenital form of DM2 has not been identified. DM2 was subsequently found to have a CCTG tetranucleotide expansion in the first intron of *Znf9* gene(Liquori et al. 2001).

The expression of RNA with the DM repeat expansions leads to the formation of RNA foci in the nucleus. It has been proposed that important RNA processing factors bind to the repeat containing RNA and are sequestered in these foci. This leads to the misregulation of RNA processing, including defects in splicing and polyadenylation. As such DM has been characterized as a RNA toxicity disease (Thornton 2014; Cho and Tapscott 2007; Turner and Hilton-Jones 2014; Machuca-Tzili, Brook, and Hilton-Jones

2005; J. E. Lee and Cooper 2009; Osborne and Thornton 2006; Douglas and Wood 2011).

There is a large body of literature describing the alterations in the RNA processing machinery (Osborne and Thornton 2006; Douglas and Wood 2011). However, the effect of the defect in RNA processing on the proteome is poorly understood. In this study, we used quantitative proteomics analysis to determine the changes in the proteomes of the skeletal muscles of DM patients.

## Materials and Methods

**Patient details:** Muscular biopsies were kindly provided by Dr. Bjarne Udd from University of Helsinki, Finland.

**iTRAQ quantitation of the proteome:** Protein extract was prepared by ultrasonicating the skeletal muscle tissue in lysis buffer (50% Trifluoroethanol 50 mM HEPES). The amountsl protein in the samples were determined using BCA assay. 10 µg of protein was aliquoted out, reduced by tris(2-carboxyethyl)phosphine, cysteine blocked by Methyl methanethiosulfonate, and digested with trypsin (1:50 :: trypsin:protein) overnight. The peptides were desalted using solid phase extraction with reverse phase microtrap coloum (Michrom Bioresources) as described in Link and La Baer. The peptides were resolublized in 7 ul 500 mM triethylammonium bucarbonate (TEAB). 85 ul of isopropyl alcohol was added to iTRAQ reagents. 12 µL of iTRAQ reagents were added to the samples, incubated with shaking for 2 hours, pooled, frozen, lyophilized, resolublized in buffer A (5% acetonitrile, 0.1 % formic acid in HPLC grade water) and stored at -80 0C. The iTRAQ labeled samples were analyzed by MudPIT essentially as

described with one change (Browne et al. 2013). The precursor ions were analyzed in the Orbitrap followed by 4 CID fragment ion scans in the ion trap to identify the peptides followed by 4 HCD fragment ion scan of the same precursors as in CID to get obtain the reporter ion intensities in the orbitrap.

**Mass spectrometry data processing and analysis**: Mass spectrometry data processing was done as described in (Samir et al. 2015). RAW files generated by LC-MS/MS experiments were searched using Sequest database search engine running under proteome Discoverer v1.4 (Thermo Scientific) to identify the peptides (Eng et al. 2008; Eng, McCormack, and Yates 1994). Sequest searches were done against an ENSEMBL database of human protein sequences. Protein assembly and reporter ion quantitation and statistical analysis were done using ProteoIQ (Premier BioSoft). $Log_2$ transformed fold change against a common control prepared from lysates of wild type myoblasts (PromoCell) was used as the measure of abundance. Correlation plot was generated in R (R Core Team 2015).

**Differential expression and pathway analysis**: One way ANOVA was used to identify differentially expressed proteins using in R using a modified version of a previously described script (Samir et al. 2015; R Core Team 2015). The modification allowed using data from control patients as covariates. Pathway analysis using the differentially expressed proteins was done using the GeneMANIA Cytoscape plugin (Montojo et al. 2010; Mostafavi et al. 2008). Cytoscape was used to visualize the network diagrams (Shannon et al. 2003). Bar graphs were generated in MS Excel.

## Results

### Proteomic analysis of myotonic dystrophy skeletal muscle biopsies.

We analyzed the proteome of skeletal muscle biopsies from 5 control subjects, 7 DM1 patients, and 6 DM2 patients. We used iTRAQ labeling followed by MudPIT analysis for the quantitation (Ross et al. 2004; Link et al. 1999). We quantitated 3575 proteins across the three groups (fig. 20A). We analyzed the list of quantitated proteins using GeneMANIA Cytoscape plugin to identify the list of overrepresented pathways in the list (Montojo et al. 2010). GeneMANIA generated network had 801 nodes and 64138 edges (fig.20B). A smaller number of nodes in the network compared to the input list of proteins represent the redundancies in the protein list due to the presence of multiple isoforms. The overrepresented pathways included a number of pathways expected to be involved in muscle physiology. The top 5 pathways enriched the list were contractile fiber part, muscle filament sliding, actin-myosin filament sliding, contractile fiber, and muscle system process (fig. 20C). In addition to muscle related pathways, energy production and translational control pathways were also enriched the list of quantitated proteins. This was expected based upon the cellular abundances of translational control proteins and metabolic enzymes.

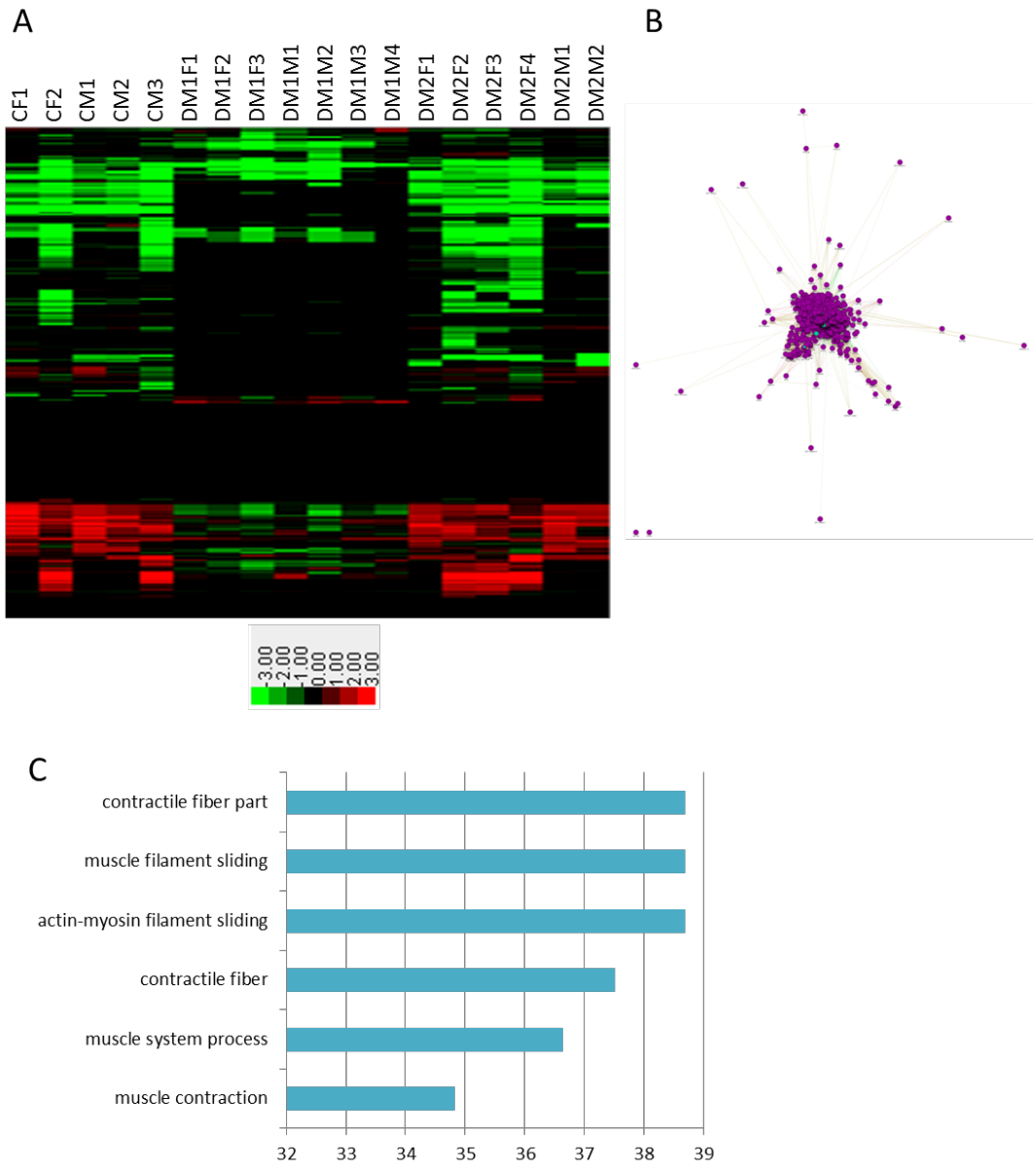



**Figure 20** – Quantitative proteomics analysis of skeletal muscles from DM patients

A) Heatmap showing the expression ratioed against the common control. B) Network interactions generated using GeneMANIA. C) Top 6 enriched pathways in the list of proteins identified across all the experiments. CF1 is control female 1, CF1 is control female 2, CM1 is control male 1, CM2 is control male 2, CM3 is control male 3, DM1F1 is DM1 patient female 1, DM1F2 is DM1 patient female 2, DM1F3 is DM1 patient female 3, DM1M1 is DM1 patient male 1, DM1M2 is DM1 patient male 2, DM1M3 is DM1 patient male 3, DM1M4 is DM1 patient male 4, DM2F1 is DM2 patient female 1, DM2F2 is DM2 patient

female 2, DM2F3 is DM2 patient female 3, DM2F4 is DM2 patient female 4, DM2M1 is DM2 patient male 1, DM2M2 is DM2 patient male 2.

Closer analysis of the GeneMANIA generated network revealed that all but two nodes were part of a subnetwork spanning rest of the nodes. The two nodes were zinc finger 788 (ZNF788) and proline rich basic protein-1 (PROB1). They were not connected to any other node in the network. ZNF788 belongs to krueppel c2h2 type zinc finger protein family (The UniProt Consortium 2015). No disease mutation in ZNF788 has been reported (Peterson et al. 2010). PROB1 has been found to be mutated in human cancers (Wu et al. 2014).

**Protein downregulated in both DM1 and DM2**

Proteins products of three genes, *RPL13A*, *P4HB*, and *MYH7B,* were found to be downregulated in both DM1 and DM2 (Figure 21A-C). RPL13A is a ribosomal protein with diverse functions in translational control (The UniProt Consortium 2015). It has not been associated with a disease in DMDM database (Peterson et al. 2010). It has been found to be mutated in human cancers, which might be reflecting its polymorphism (Wu et al. 2014). RPL13A was found to have extraribosomal function in translational control of interferon regulated mRNAs. This led to the proposition of the depot hypothesis (Mazumder et al. 2003; Ray, Arif, and Fox 2007). According to ENSEMBL database (Ensembl release 82) *RPL13A* pre-mRNA contains 8 exons and 7 introns (Cunningham et al. 2015). A number of transcripts with retained introns have been reported. These transcripts are not translated into proteins (Cunningham et al. 2015). Since aberrant splicing is a common defect in DM1 and DM2, it provides a mechanism by which RPL13A protein levels might be downregulated in both DM1 and DM2.

GeneMANIA based pathway analysis to find pathways that might be affected by changes in RPL13A levels revealed translational control as the main pathway. Most of the proteins in the resulting network were ribosomal proteins (Fig. 21A). One of the members of the RPL13A network is Mago-Nashi Homolog (MAGOH), which is a member of exon junction complex (The UniProt Consortium 2015). MAGOH regulates neural stem cell division. A haploinsufficiency in *Magoh* leads to reduced brain size in mouse (Silver et al. 2010).

Second protein downregulated in both DM1 and DM2 is prolyl 4-hydroxylase, beta polypeptide (P4HB) (Fig. 21B). P4HB belongs to protein disulfide isomerase family of proteins (The UniProt Consortium 2015). P4HB has not been found to be associated with a human disease, but it has been found to be mutated in human cancers that might be reflecting its polymorphism (Peterson et al. 2010; Wu et al. 2014). P4HB pre-mRNA contains 11 exons and 10 introns. Similar to RPL13A, a number of transcripts with retained introns have been described. These transcripts are not translated into proteins suggesting a mechanism of downregulation dependent upon aberrant splicing (Cunningham et al. 2015). GeneMANIA analysis of a network generated from P4HB revealed association with metabolic enzymes (Fig. 21B).

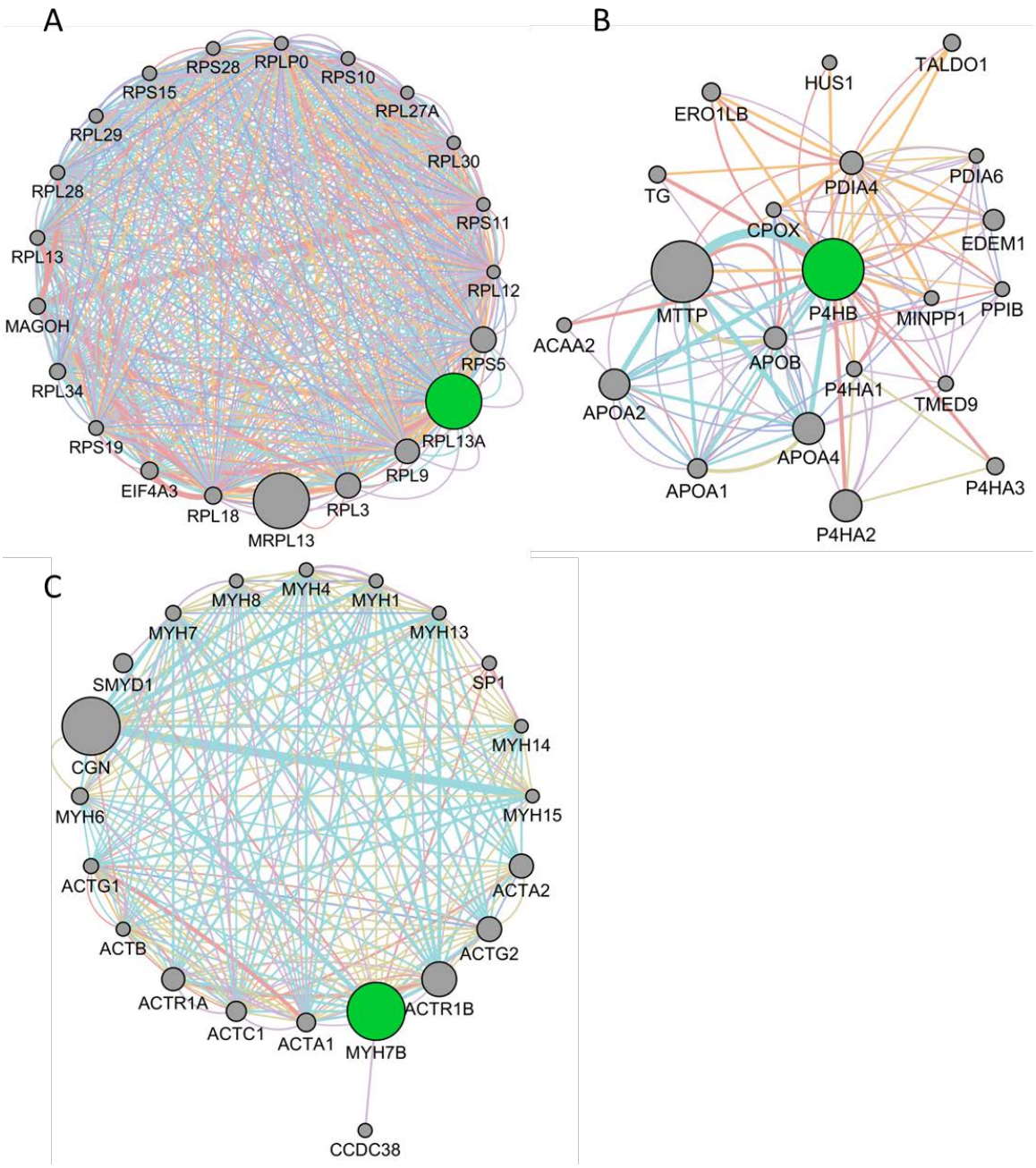



**Figure 21** – Network of interactions of the three genes downregulated in both DM1 and DM2

Networks were generated using GeneMANIA Cytoscape plugin. The gene corresponding to the identified protein is colored green. Grey nodes were inferred from by GeneMANIA. A) Network of interactions of ribosomal protein *RPL13A*. B) Network of interactions of *P4HB*. C) Network of interactions of *MYH7B*.

The third and final protein downregulated in both DM1 and DM2 was Myosin heavy chain 7B (MYH7B) (Figure 21C). MYH7B has 9 transcripts that are generated by alternative splicing (Cunningham et al. 2015). The largest protein coding isoform has 43 exons and 42 introns. The smallest protein coding isoform has 4 exons and 3 introns. There is a transcript with retained intron that does not code for a protein (Cunningham et al. 2015). MYh7B has been associated with left ventricular noncompaction disease (Cunningham et al. 2015). The large numbers of introns in *MYH7B* pre-mRNA makes it very susceptible to defects in splicing machinery. GeneMANIA analysis of a network generated from MYH7B revealed interactions with other myosins as well as other cytoskeletal components (Fig. 21C).

**Protein upregulated in both DM1 and DM2**

Protein products of three genes were found to upregulated in both DM1 and DM2 (Fig. 22A-C). The first gene is ATPase, Ca++ transporting, cardiac muscle, fast twitch 1 (ATP2A1). ATP2A1 is an ATP dependent calcium ion transporter responsible for reuptake of Ca++ ions into the sarcoplasmic reticulum in striated muscles. A mutation in ATP2A1 has been found to be associated with Brody disease (Odermatt et al. 2000). Brody disease is a rare inherited myopathy characterized by delayed skeletal muscle relaxation and silent cramps (Voermans et al. 2012). This is similar to myotonia observed in DM, in which patients have difficulty relaxing their muscles. Brody disease is associated with a loss of Calcium uptake function of ATP2A1. The significance of an upregulation of ATP2A1 in both DM1 and DM2 is not immediately clear. GeneMANIA analysis of the network generated from ATP2A1 revealed its associations with a number of cytoskeleton components (Fig. 22A).

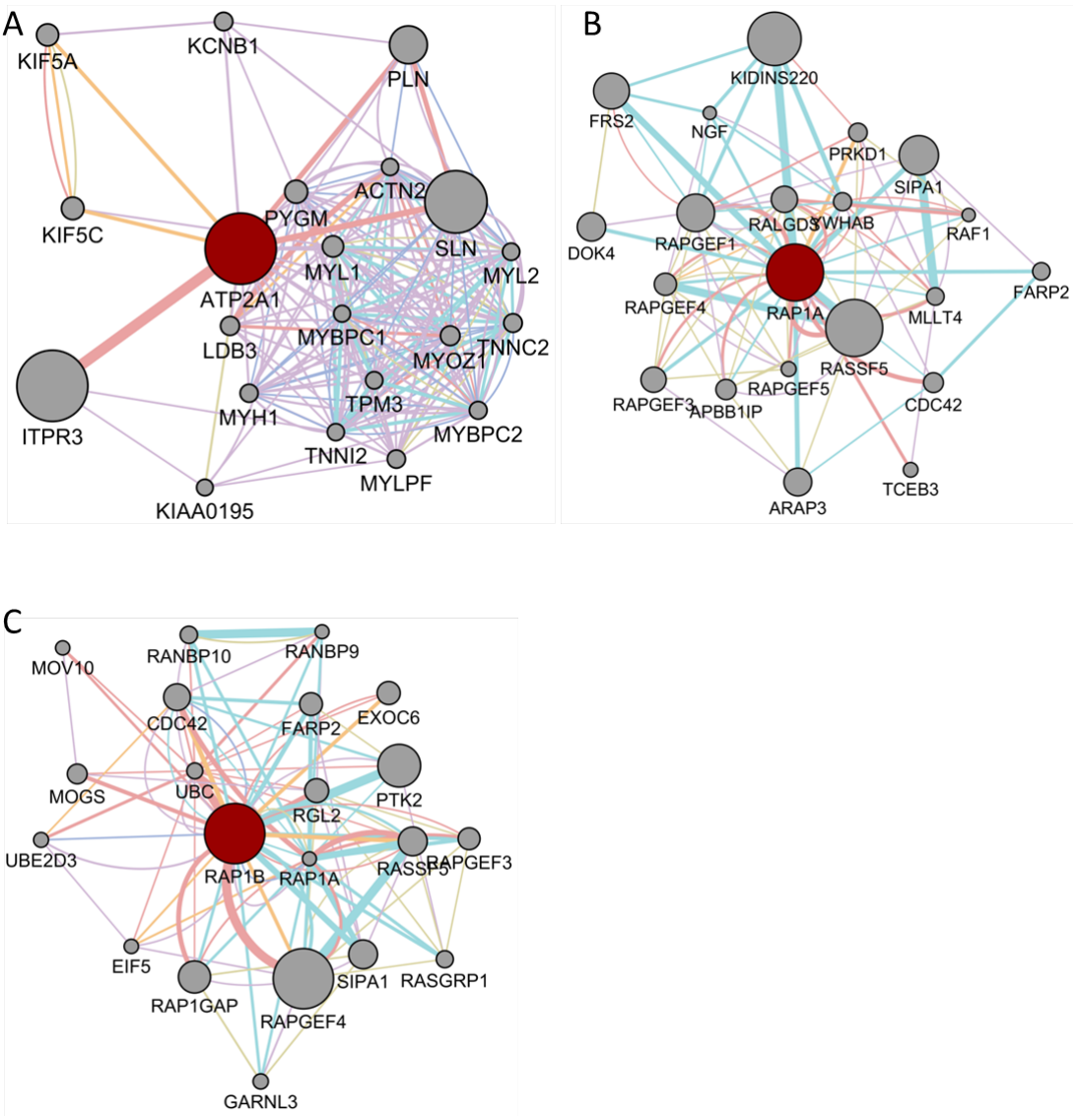



**Figure 22** – Network of interactions of the three genes upregulated in both DM1 and DM2

Networks were generated using GeneMANIA Cytoscape plugin. The gene corresponding to the identified protein is colored red. Grey nodes were inferred from by GeneMANIA. A) Network of interactions of ribosomal protein *ATP2A1*. B) Network of interactions of *RAP1A*. C) Network of interactions of *RAP1B*.

The other two proteins upregulated are Ras-related protein Rap-1A (RAP1A) and Ras-related protein Rap-1b (RAP1B) (Fig. 22B-C). They are paralogous proteins with similar sequences. They have been reported because it was not possible to distinguish

between the two based upon the identified proteins. They are members of RAS oncogene family. RAP1A has 4 transcripts generated from alternative splicing, while RAP1B has 33 transcripts. Neither protein has been found to be associated with an inherited human disease (Peterson et al. 2010). GeneMANIA network generation and analysis using RAP1A and RAP1B individually revealed associations with signaling pathways as well as translation initiation factor. RAP1B was found to associate with translation initiation factor eIF5 as well ubiquitin.

**Proteins downregulated in DM1**

GeneMANIA network generation and pathway analysis revealed muscle function related pathways to be overrepresented in the list of proteins downregulated only in DM1 (Fig. 23A-B). There are 56 downregulated gene products in the GeneMANIA network (Fig. 23A). The downregulated proteins included critical components of muscles including Titin (TTN), myosin light chains, and myosin heavy chain. It also included energy producing metabolic enzymes such as Phosphoglycerate kinase 1 (PGK1) and Fructose-bisphosphate aldolase A (ALDOA). Both of the proteins have been associated with neuromuscular diseases (The UniProt Consortium 2015). Top 5 misregulated pathways in the list of proteins downregulated in DM1 are muscle filament sliding, actin-myosin filament sliding, actin-mediated cell contraction, actin filament-based movement, and muscle contraction (Fig. 23B).

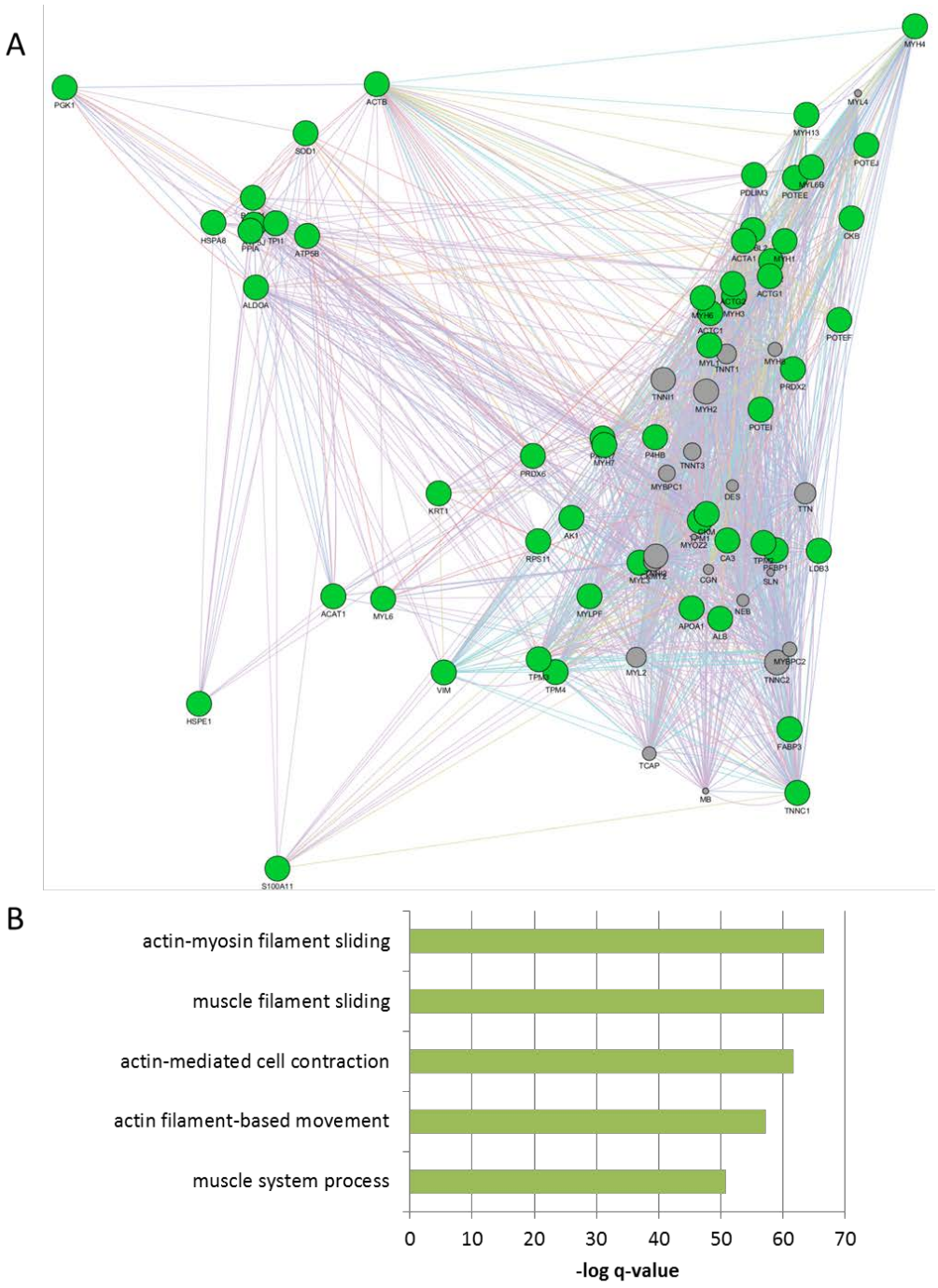



**Figure 23** – Pathway analysis of proteins downregulated in DM1

Networks were generated using GeneMANIA Cytoscape plugin. The nodes corresponding to the identified proteins are colored green. Grey nodes were inferred from by GeneMANIA. A) The network interactions generated using GeneMANIA. B) Top 5 enriched pathways in the list.

Another protein downregulated in DM1 is Superoxide dismutase (SOD1). A number of mutations in SOD1 has been found to be associated with inherited familial amyotrophic lateral sclerosis (ALS) (Nakano et al. 1994; Rosen et al. 1993; Kostrzewa, Burck-Lehmann, and Müller 1994; P. M. Andersen et al. 2003). ALS shares many similarities with DM (Robberecht and Philips 2013). Misregulation of SOD1 is an interesting candidate event that could explain the similarities between ALS and DM.

Heat shock 70kDa protein 8 (HSPA8) is also downregulated in DM1. HSPA8 is a multifunctional protein involved in activation of transcription, protein quality control and bacterial lipopolysaccharide response in immune cells (Matsumura, Sakai, and Skach 2013; Triantafilou, Triantafilou, and Dedrick 2001; Yahata et al. 2000). Although HSPA8 has not been implicated in inherited human diseases, as part of PRP19-CDC5L complex it binds to all the core components of spliceosomes (Makarova et al. 2004). A downregulation of HSPA8 might contribute to the splicing anomalies observed in DM1.

**Proteins upregulated in DM1**

The analysis of proteins upregulated in DM1 revealed protein products of 39 genes (Fig. 24). The network generated in GeneMANIA contains the 39 query genes and 20 associated genes (Fig. 24A). Top five enriched pathways in the list are positive regulation of mitochondrial membrane permeability involved in apoptotic process, mitochondrial outer membrane permeabilization, protein insertion into mitochondrial membrane involved in apoptotic signaling pathway, regulation of mitochondrial outer membrane permeabilization involved in apoptotic signaling pathway, and regulation of protein insertion into mitochondrial membrane involved in apoptotic signaling pathway (Fig. 24B).

One protein upregulated in DM1 is Elongation factor 1-alpha 2 (EEF1A2). EEF1A2 is a translation elongation factor belonging to TRAFAC class translation factor GTPase superfamily and EF-Tu/EF-1A subfamily. It contains one tr-type G (guanine nucleotide-binding) domain (The UniProt Consortium 2015). EEF1A2 has two transcripts generated from alternative splicing. Both of the transcripts code for a 463 amino acid protein. One of the mRNAs is made 8 exons, 7 of which constitute the coding region while the other consists of 7 exons all of which are coding (Cunningham et al. 2015). EEF1A2 has been found to be associated with Epileptic encephalopathy, early infantile, 33 (Veeramah et al. 2013; de Ligt et al. 2012). EEF1A2 has also recently been found to be associated with Mental retardation, autosomal dominant 38 (Nakajima et al. 2015). Given the neurological symptoms of DM1, the aberrant function EEF1A2 in DM1 seems to be one of the contributing factors in DM1 (de León and Cisneros 2008).

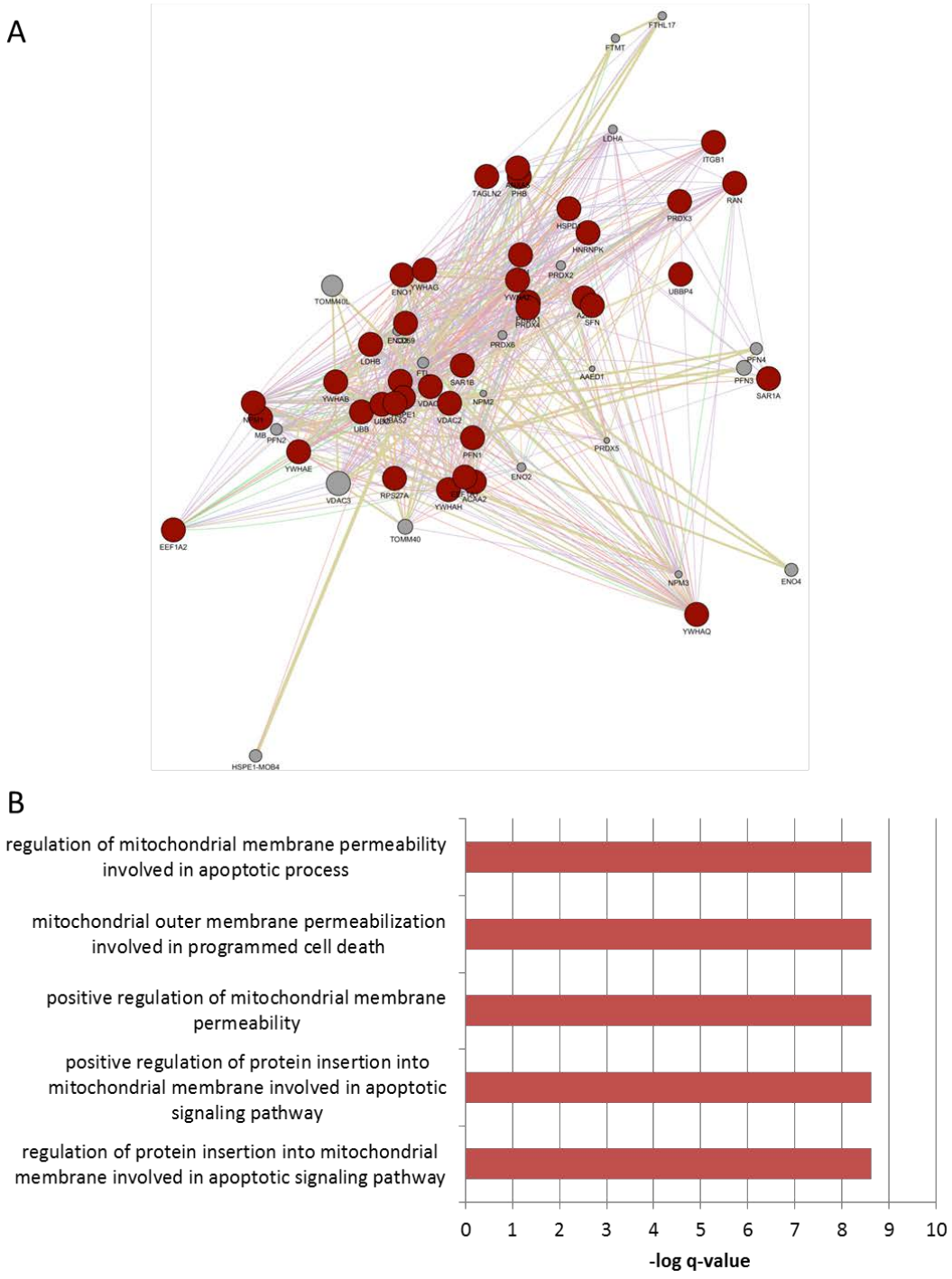



**Figure 24** – Pathway analysis of proteins upregulated in DM1

Networks were generated using GeneMANIA Cytoscape plugin. The nodes corresponding to the identified proteins are colored red. Grey nodes were inferred from by GeneMANIA. A) The network interactions generated using GeneMANIA. B) Top 5 enriched pathways in the list.

Another translational control protein upregulated in DM1 is ribosomal protein S27a (RPS27A). RPS27A contains ubiquitin on its N-Terminus. It has 10 transcripts, 4 of which are protein coding (Cunningham et al. 2015). The largest protein coding transcript consists of 6 exons, 5 which of which are coding exons. The smallest protein coding transcript consists of 5 exons with 4 coding exons. Six non-coding transcripts generated from *RPS27A* gene contain retained introns (Cunningham et al. 2015). RPS27A has not been found to associated with an inherited human disease (Peterson et al. 2010). RPS27A is one of the major contributors to the ubiquitin pool in the cells (Bianchi et al. 2015). It is also involved in regulation of p53 level through degradation of its regulator MDM2. Overexpression of RPS27A was shown to stabilize and increase the amount of p53 (Xiong et al. 2011). It has been shown that DM1 muscles undergo an increased rate of apoptosis (Loro et al. 2010). An increase in the amount of of RPS27A, which might lead to an increase in p53 and apoptosis, provides a putative mechanism for explaining this observation.

## Discussion

A common theme in the proteins whose expression is altered in both DM1 and DM2 is the presence of alternative splicing in their mRNA maturation. This is in agreement with the previous studies that showed a central role for aberrant splicing in DM pathogenesis. In addition to supporting the previous studies, we have identified a number of candidate proteins pathways that are attractive targets for follow up studies.

Formation of cataracts is a common symptom in both DM1 and DM2. This was initially thought to be caused by a decrease in expression of *SIX5* gene in DM1 (Klesert et al. 1997; Klesert et al. 2000; Sarkar et al. 2000). The decrease in *SIX5* expression is attributed to the changes in the heterochromatin around the *DMPK* locus. Since *ZNF9* locus is not on the same chromosome, this suggested that the decrease in SIX5 protein may not be the common mechanism behind cataract formation. It was also noted that the type of cataracts observed in mouse deficient in *Six5* and DM patients were different (Rhodes et al. 2012). To identify candidate causal agents, a microarray study was done on lens from DM patients (Rhodes et al. 2012). The study revealed differential expression of interferon responsive genes. These genes might be activated in response to double stranded RNA (dsRNA) in the cells (Rhodes et al. 2012). An important feature of immune response is the temporal regulation of gene expression so that the probability of a runaway immune response leading death of the host is reduced (R. Mukhopadhyay et al. 2009).

A critical player in regulating the inflammatory response is the GAIT complex (R. Mukhopadhyay et al. 2009). GAIT complex is heterotetrameric complex consisting of IFN-γ-activated inhibitor of translation (GAIT) complex comprising glutamyl-prolyl tRNA synthetase (EPRS), NS1-associated protein 1 (NSAP1), RPL13A, and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). GAIT complex is maintained in an inactive state in which RPL13A is not present. RPL13A phosphorylation on the ribosome triggers its release from the ribosome. Its subsequent binding to the inactive GAIT complex activates it. The activated GAIT complex binds to the 3'UTR of target mRNAs and inhibits translation initiation (R. Mukhopadhyay et al. 2009). A decrease in RPL13A

protein levels can cause failure of activation of GAIT complex. This might lead to a decrease in the resolution of inflammation and can explain the observation of differential expression of immune responsive genes in DM eyes (Rhodes et al. 2012).

A number of other candidate genes have been identified in this study that can contribute to the different symptoms observed, especially in DM1. Surprisingly, no DM2 specific changes in the protein levels were observed. This could be because of the milder phenotype of DM2. Since we solely relied on the statistical significance obtained from our linear regression analysis, a technical reason could be the high variance in the proteomic expression patterns of DM2 patients. Nevertheless, this study provides an important resource and catalog of proteins differentially expressed in DM1 and DM2.

Studies in mouse models of DM2 had suggested a key requirement for *ZNF9* protein, whose loss was able to recapitulate some of the symptoms of DM2 (W. Chen et al. 2007). This suggested that loss of function of a protein was at least contributing to DM pathogenesis in DM2. It became imperative to find the molecular function of *ZNF9* protein. Previous studies in the Link lab had suggested a role for translational control in DM2 pathogenesis due to the molecular function of *ZNF9* protein. *ZNF9* was found to be involved in IRES mediated translation in yeast as well as cell culture models of DM1 (Gerbasi and Link 2007; Sammons et al. 2010). Another common theme in this study was the misregulation of many translational control proteins. It provides further evidence for the critical role of translational control in health and disease.

# Chapter V

## Conclusions and Future Directions

In my graduate research, I studied three problems of the regulation of biochemical state and the information repertoire of cells and tissues. The first problem involved studying the regulation of biochemical states by the information from outside the cellular boundary. I built a conceptual basis for interpreting complex cellular responses to multiple concurrent environmental stimuli. The second problem involved testing the ribosome filter hypothesis. A ribosome filter is a ribosome mediated regulatory element that controls the amount of information flow from specific mRNA transcripts to proteins. The third and final problem was to investigate the proteomic changes in the skeletal muscles of myotonic dystrophy patients. These proteomic changes might have been caused by the disruption of information flow due to microsatellite repeat expansions in the genomes of the patients.

### Cellular responses to environmental stimuli

The complement of molecules contained in a cell, including the cell surface, constitutes its biochemical state. The biochemical state is a reservoir of information. The genetic material of an organism, its DNA or RNA, contains the template information which is used to synthesize all of the necessary molecules, or in some cases the molecules that can make or modify those molecules. The flow of information from the genetic material to the functional molecules, for example RNAs or Proteins, is known as the central dogma (Crick 1970). An organism uses the information flow to respond to and modify its environment. This flow of information, however, is not linear. The DNA bases can be covalently modified that affects the information flow through transcription

(Suzuki and Bird 2008; Robertson 2005; T. Phillips 2008). Similarly, RNAs have been found to be extensively edited or modified, which also modifies their information content (Nishikura 2010). In most of these processes, the class of molecules responsible for the function, the molecular actuators, are the proteins. Although the backbone of the protein can be synthesized solely based upon the information content in the genetic material, with the notable exception of those synthesized from edited mRNA templates, they themselves can be covalently modified after synthesis. The covalent modifications can have many functions that include altering the biochemical activity of the proteins, modulating sub-cellular localization or binding to cofactors, and targeting or protecting for degradation (Wells, Whelan, and Hart 2003; P. Anderson and Kedersha 2009; Wold 1981; Lodish 1981; Nussinov et al. 2012; Vucic, Dixit, and Wertz 2011; Beltrao et al. 2013; Terman and Kashina 2013). This adds another layer of information content. There are two fold consequences of this added layer: (1) it allows the living beings to store more information (increase in information repertoire) that it can use to respond to a wider range of environments, and (2) the prior biochemical state, including its complement of proteins and their covalent modifications, determines the exact cellular response.

The information repertoire of the cells, therefore, is dependent upon two factors; (1) the information in the genetic material, and (2) the information in the environmental stimuli, including the prior stimuli they had been exposed to. Extrapolating this logic, since genes and stimuli are just packets of information that cells use for their continued survival, as an abstraction they can be thought to be the same. There is a very

important consequence of this assumption, viz. the tools used for studying genes can be applied for stimuli.

The information flow from genes has been a well-studied problem. Consequently, there is a large body of work and a well-defined conceptual basis associated with it. The conceptual basis for studying information flow and integration of information from multiple genes is called gene interactions. It provided, in many cases, an easy interpretation of observed changes in the characteristics or traits upon alterations in genes. It has also been used to decipher the order in biochemical and signaling pathways making it a very valuable tool for research (St Johnston 2002). With the abstraction that stimuli are analogous to genes, most of these concepts and tools can become available for studying the effects of multiple concurrent stimuli.

I used one of the abstractions, the dominance in gene interactions, as a tool to identify proteins and transcripts that are important for responding to specific stimulus in *S. cerevisiae*. The assumption here was that if a protein or transcript contains the information important for responding to a stimulus, the effect of the stimulus would be dominant over the effects of an unrelated stimulus. There is a caveat associated with this approach that makes using absolute dominance fraught with false negative results. If a biomolecule contains information for responding to multiple stimuli among the set under investigation, dominance will not observed. This would lead to a false negative result. To address this caveat and further refine the idea follow up studies are needed.

**Follow up studies about environmental interactions and epistasis**

The follow up studies on environmental interactions and epistasis can be classified into two categories – (1) showing the general applicability of the ideas, and (2) understanding the mechanistic basis behind the interactions.

The validity of the concepts of environmental interactions and epistasis has been shown only in *S. cerevisiae* with the transcriptomic and proteomic changes (Samir et al. 2015). Paucity of published datasets in other organisms makes it difficult to test the general validity of the concept. There is a published dataset in which liquid cultures of *A. thaliana* cells were used to study the effects of carbon dioxide concentrations and high salinity on the transcriptome as well as the metabolome (Kanani, Dutta, and Klapa 2010; Dutta et al. 2009). This dataset is ready resource to test the validity of the conceptual basis in plants. The post-translational modifications data in the study of high osmolarity or pheromone signaling provides a resource for testing the hypothesis with a different cellular response, albeit still in *S. cerevisiae* (Vaga et al. 2014).

HeLa cells could be used to test the hypothesis in mammalian systems. The stimuli used could be increasing concentrations of two inhibitors that target different signaling pathways. Rapamycin can be used to inhibit the mTOR pathway while one of the “inhibitors of Wnt response” compounds (Law 2005; Baozhi Chen et al. 2009). Transcriptomic responses can be measured by RNA-Seq and the proteomic responses can be measured by iTRAQ labeling liquid chromatography tandem mass spectrometry (Ross et al. 2004; Cloonan et al. 2008; Lister et al. 2008; Nagalakshmi et al. 2008; B. T. Wilhelm et al. 2008; Mortazavi et al. 2008). Phosphoproteomics analysis can be used to assay the changes in phosphorylation states of the target proteins (Winter et al. 2012).

Mouse models of T-cell and B-cell activation can be used to test the validity in mammalian systems *in vivo*. An ovalbumin specific transgenic mouse, OT-1 can be used as the model system. This strain of mice contains T-cell receptor that is specific to a peptide antigen generated from ovalbumin (Hogquist et al. 1994; Clarke et al. 2000). The stimuli used can be different concentrations of the antigenic peptide and the adjuvants. The transcriptomic and proteomic responses can be assayed in peripheral blood mononuclear cells (PBMC) and polymorphonuclear cells (PMN). Metabolomic changes can be assayed in blood plasma. An important feature of this study design is the use of different concentrations of the stimuli. A multiple linear regression model built with the different concentrations can help identify the molecules that are important for responding to multiple stimuli if their response is dose dependent.

The second aspect, to identify a mechanistic basis for the observed phenomenon can be done in *S. cerevisiae* using the same stimuli as before with quantitative proteomics analysis. In this study, a time course experiment would be needed. The order of application of the second stimuli would need to be changed. This experiment design would allow assay of kinetics of modulation of the proteome. The biomolecular complexes, or their components, that are differentially regulated in the kinetics experiments can be the candidates, with an assumption that they are involved in the modulation information flow, for pursuing more in depth biochemical and genetic analyses.

**Regulation of proteome by the ribosome filter**

In this study, I used two complementary techniques, quantitative mass spectrometry and cryo-EM (in collaboration with the Joachim Frank lab), to test the

ribosome filter hypothesis by measuring changes in the protein composition of the ribosomes. The ribosome filter provides a mechanism for the ribosomes to regulate information flow through translational control. It can act through the use of specific paralogs for translating specific mRNAs more efficiently.

I identified 11 ribosomal proteins whose abundances in the purified ribosomes were changing. The list included a paralog pair, Rpl8a and Rpl8b. Polysome profiling with the null mutants of either Rpl8a or Rpl8b suggested that their functions are not redundant. I identified 80S ribosomes with substoichiometric protein compositions using cryo-EM. A time course experiment after shifting cells from a glucose containing media to a glycerol containing media followed by ribosome purifications cryo-EM analysis showed that the proportions of the substoichiometric ribosomes were changing. This suggested the cells dynamically regulate their ribosome composition.

In conclusion, I have used two techniques to assay the changing composition of ribosomes. These changes support the ribosome filter hypothesis. Follow up experiments are needed to identify the transcripts affected by Rpl8a/b mediated ribosome filter as well the underlying mechanism.

**Follow up studies on ribosome mediated translational control**

I have identified a candidate paralog pair, *Rpl8a* and *Rpl8b*, which might have specific functions in translation. They might be required for translation of specific transcripts. However, the identity of the transcripts is not known. Ribosome footprint profiling after RNAse treatment can be used to identify such transcripts. Once the transcripts have been identified, two complementary approaches can be used to

decipher the mechanism through which RPL8A or RPL8B helps translate specific transcripts. In the first approach, a bioinformatics search can be performed to find sequence motifs overrepresented in transcripts that need a specific paralog. Reporter assays can be done to check the effects of the sequence motifs on translation. In the second approach, *in vitro* translation reactions using the ribosomes from the null mutants can be used to directly assay the rate of translations of the identified transcripts. The transcripts will need to be *in vitro* transcribed to ensure quality control across experiments.

After testing the ribosome filter hypothesis in yeast, it could be tested in mammalian cell culture system using a similar approach. In this case, HeLa cells can be used as the model system because they are one of the most well characterized mammalian cell culture systems.

Another question that arises from this study deals with the exact mechanism(s) of change in the composition. There can be three models that explain the changes in the composition. In the first model, there are free floating ribosomal protein paralogs in the cytoplasm or nucleus. Upon specific signaling cues, the paralog on the ribosome is exchanged for the free floating one. The advantage of this model for the cell is that the kinetics of changing composition will be the fastest. However, the cells would need to synthesize proteins that they do not need at any given time. This will mean expenditure of energy to keep the system primed. Since translation is the most energetically costly process in the cells, and translation of ribosomal proteins constitutes a very large chunk of the total expenditure, the energetic cost for the cell will be very high. In the second model, new ribosomes are synthesized with specific paralogs in response to signaling

cues. In this model, there is less expenditure of energy as cells synthesize only the ribosomal proteins and rRNAs that they need at any given time. However, a disadvantage of this model for the cell is that the response time might be very high. This is because the biogenesis of ribosomes requires multiple steps. In the third model, the cells degrade the ribosomes that contain the paralogs that they do not need. The short term energy cost of this model is the least because it does not require synthesis of new ribosomes. However, the long-term energy cost of this model might be the highest because it does not involve synthesis of ribosomes only with the paralog that is needed. This means the cells will have to continuously spend energy on making ribosomes that it does not need and then degrade it. To decipher the exact mechanism, labeling with stable isotopes followed by mass spectrometry quantitation can be used.

**Regulation of proteome by RNA repeat expression in myotonic dystrophy**

RNA repeat expansion disrupts the information flow by sequestering important RNA binding factors that regulate the transfer and modification of information from genome to proteins. I have identified several candidate proteins that might have roles in DM pathogenesis. This included RPL13A, P4HB, and MYH7B. This study is a starting point for further studies with these candidate proteins to dissect the mechanism of disruption in information repertoire that leads to DM.

**Follow up studies about the proteomic changes in myotonic dystrophy**

A number of interesting candidate proteins were identified that might play roles in DM pathogenesis. Since I am interested in translational control, especially the regulation by ribosomes, I think *RPL13A* is a very interesting candidate. A previous transcriptomic study had identified misregulation of inflammatory response genes in eye

lenses of DM patients. Since RPL13A protein has been shown to regulate inflammation in a temporal manner, it might be playing a role. Inflammation inside the eye has been found to be associated with cataract formation (Hodge, Whitcher, and Satariano 1994; Durrani et al. 2004). Since RPL13A inhibits translation of proinflammatory proteins as part of GAIT complex, its loss of function can cause persistent inflammation. This could explain the observation of cataracts in DM patients. The function of RPL13A can be studied in mouse models.

Since *ZNF9* has been found to be involved in IRES mediated translation, it would be informative to find its *in vivo* targets. RNA-pulldown experiments followed by RNA-Seq can be used to identify its targets. An attractive alternative is the PAR-CLIP for identification of RNA binding sites for candidate RNA binding proteins on the transcripts. An unrelated RNA binding protein, such as PABP can be used as a control in this experiment. Once the targets have been identified, their regulation in DM2 patients can be studied. Yeast, cell culture and mouse models can be used to study the effect of their loss of function. X-Ray crystallography can be used to determine the structural basis for Znf9 binding to mRNAs.

## Appendix A – Table 1: Proteins Quantitated in Environmental Interactions Study

https://drive.google.com/open?id=0BxmfH2AgA_HkQ2NnTTZ5eGZQa3M

## Appendix B – Table 2: GeneMANIA pathway analysis output for HT stimulus

https://drive.google.com/open?id=0BxmfH2AgA_HkRTZNZGROeTE1Z1k

## Appendix C – Table 3: GeneMANIA pathway analysis output for G stimulus

https://drive.google.com/open?id=0BxmfH2AgA_HkQ3h3WFFVcF9vZ2c

## Appendix D – Table 4: GeneMANIA pathway analysis output for HT+G stimulus

https://drive.google.com/open?id=0BxmfH2AgA_HkbzFZMmRSRGJEYmc

## Appendix E – Table 5: GeneMANIA pathway analysis output for HT stimulus dominance

https://drive.google.com/open?id=0BxmfH2AgA_Hkd2VEd0tkdHUtT28

## Appendix F – Table 6: GeneMANIA pathway analysis output for G stimulus dominance

https://drive.google.com/open?id=0BxmfH2AgA_HkQnE3M1ZwalpwT0E

## Appendix G – Table 7: GeneMANIA pathway analysis output for non-specific environmental response in protein expression

https://drive.google.com/open?id=0BxmfH2AgA_HkcGc4bXhSOVF3V2c

## Appendix H - Table 8: GeneMANIA pathway analysis output for discordance in protein expression

https://drive.google.com/open?id=0BxmfH2AgA_HkVFdiMW4xbGdaQzA

## Appendix I – Table 9: GeneMANIA pathway analysis output for suppression in protein expression

https://drive.google.com/open?id=0BxmfH2AgA_HkeHJ4bWJvTzBlSEE

## Appendix J – Table 10: GeneMANIA pathway analysis output for environmental epistasis in protein expression

https://drive.google.com/open?id=0BxmfH2AgA_HkRTI2aWdZNUsyWDA

## Appendix K – Table 11: GeneMANIA pathway analysis output for no environmental epistasis in protein expression

https://drive.google.com/open?id=0BxmfH2AgA_HkOEhOUTZiUGpRaW8

## Appendix L – Table 12: Complete data matrix of transcripts

https://drive.google.com/open?id=0BxmfH2AgA_HkMTF5dzdwM1NLUm8

## Appendix M – Table 13: GeneMANIA pathway analysis output for environmental epistasis in transcript expression

https://drive.google.com/open?id=0BxmfH2AgA_HkaGJiUXNLT0VjN2c

## Appendix N – Table 14: GeneMANIA pathway analysis output for dominance of NS

https://drive.google.com/open?id=0BxmfH2AgA_Hkbzg0Q1lMYy1QVms

## Appendix O – Table 15: GeneMANIA pathway analysis output for dominance of AN

https://drive.google.com/open?id=0BxmfH2AgA_HkUWdSbm5KclF5YWM

## Appendix P – Table 16: Doubling times under the 8 growth conditions

https://drive.google.com/open?id=0BxmfH2AgA_HkbkdCVGp2WWJQMlk

## Appendix Q – *R* source codes for the analysis in environmental interactions analysis

https://drive.google.com/open?id=0BxmfH2AgA_Hkdmp1ZG1STllYZkU

https://drive.google.com/open?id=0BxmfH2AgA_HkZlp0OGNJTU16N0k

https://drive.google.com/open?id=0BxmfH2AgA_HkMUMySWxOWmdNSXc

https://drive.google.com/open?id=0BxmfH2AgA_HkMHozNjZ4NDA5OUU

## Appendix R – *Python* source code for GoZilla

https://drive.google.com/open?id=0BxmfH2AgA_HkMlpiNTBQSUVZMkU

https://drive.google.com/open?id=0BxmfH2AgA_HkOTRGSVdsOWM0LVU

https://drive.google.com/open?id=0BxmfH2AgA_HkZ1h4bkxMellCUTg

https://drive.google.com/open?id=0BxmfH2AgA_HkemJCX2puNmYtSW8

## Appendix S – *Python* source code for CompZilla

https://drive.google.com/open?id=0BxmfH2AgA_HkVnF2WkVKdy1PWUU

https://drive.google.com/open?id=0BxmfH2AgA_HkanRTLWdDeEUyXzA

https://drive.google.com/open?id=0BxmfH2AgA_HkZUZ6SDM2LWxwWmM

## Appendix T – Table 17: Purified ribosomes quantitative proteomics dataset.

https://drive.google.com/open?id=0BxmfH2AgA_HkOHZacmI0WFcwcFU

## Appendix U – Table 18: Myotonic dystrophy quantitative proteomics datasets.

https://drive.google.com/open?id=0BxmfH2AgA_HkRUwwci05U0FFaGs

# Appendix V – Manuscript – 1: Environmental Interactions and Epistasis Are Revealed in the Proteomic Responses to Complex Stimuli

PLOS | ONE

RESEARCH ARTICLE

# Environmental Interactions and Epistasis Are Revealed in the Proteomic Responses to Complex Stimuli

**Parimal Samir[1], Rahul[2], James C. Slaughter[3], Andrew J. Link[1,4,5]***

**1** Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **2** Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada, **3** Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **4** Department of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America, **5** Department of Chemistry, Vanderbilt University, Nashville, Tennessee, United States of America

* andrew.link@vanderbilt.edu



OPEN ACCESS

**Citation:** Samir P, Rahul, Slaughter JC, Link AJ (2015) Environmental Interactions and Epistasis Are Revealed in the Proteomic Responses to Complex Stimuli. PLoS ONE 10(8): e0134099. doi:10.1371/journal.pone.0134099

**Editor:** Ben Lehner, CRG, SPAIN

**Received:** May 19, 2015

**Accepted:** June 26, 2015

**Published:** August 6, 2015



**Data Availability Statement:** Data have been deposited to Proteome Exchange: PXD002371.

**Funding:** AJL and PS were supported by NIH grant GM064779 and Vanderbilt University School of Medicine IDEAS Program grant 1-04-066-9530 to AJL.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

Ultimately, the genotype of a cell and its interaction with the environment determine the cell's biochemical state. While the cell's response to a single stimulus has been studied extensively, a conceptual framework to model the effect of multiple environmental stimuli applied concurrently is not as well developed. In this study, we developed the concepts of environmental interactions and epistasis to explain the responses of the *S. cerevisiae* proteome to simultaneous environmental stimuli. We hypothesize that, as an abstraction, environmental stimuli can be treated as analogous to genetic elements. This would allow modeling of the effects of multiple stimuli using the concepts and tools developed for studying gene interactions. Mirroring gene interactions, our results show that environmental interactions play a critical role in determining the state of the proteome. We show that individual and complex environmental stimuli behave similarly to genetic elements in regulating the cellular responses to stimuli, including the phenomena of dominance and suppression. Interestingly, we observed that the effect of a stimulus on a protein is dominant over other stimuli if the response to the stimulus involves the protein. Using publicly available transcriptomic data, we find that environmental interactions and epistasis regulate transcriptomic responses as well.

## Introduction

In their native environments, cells continuously respond to a complexity of environmental stimuli. These include ambient temperature fluctuations, nutrient availability, signaling molecules, and physical forces. In response, cells adjust their biochemical state through multiple mechanisms including the differential production, modification, and degradation of transcripts and proteins [1,2,3,4,5]. Both extracellular signaling and the metabolic environment strongly influence a cell's growth and responses to therapeutic treatments [6,7,8,9]. Model organisms

have been used extensively to study cellular responses to individual and combinations of environmental stimuli [1,10,11,12,13,14,15,16,17]. We extend these approaches by developing and testing a novel conceptual framework to study proteomic responses of cells to the combinatorial effects of multiple concurrent environmental factors. We have modeled our analysis of these complex environmental interactions using the concepts of gene interaction and genetic epistasis.

Gene interaction is defined as the interaction between genes at different loci that affect the same characteristic or a trait [18]. Classically, genetic epistasis is referred to a type of gene interaction in which a mutation at one locus masks or suppresses the phenotype of a mutation at a different locus [18,19]. To test the independence of the effects of individual genes, genetic epistasis has also been defined mathematically as a type of gene interaction in which the combined effect of two or more mutations is not the sum of the effects of the individual mutations [20,21,22].

Conceptually, the problem of studying multiple concurrent environmental stimuli is similar to the problem of studying the effects of multiple genetic mutations. The product of a gene functions as part of one or more functional modules in concert with the products of many genes. The changes in a gene, for example its loss of function or gain of function, affects the phenotype due to the changes in the activity of the functional modules. If multiple genetic alterations are present, the total effect is due to the integration of the effects of the individual alterations through the functional modules. Similarly, environmental stimuli affect the biochemical state of the cells through specific sensing, signaling, and response modules. Concurrent application of multiple environmental stimuli, similar to the genetic alterations, requires the integration of information from these modules to mount an optimal response. By considering an environmental stimulus as an analogue of a gene, we hypothesized that the concepts of gene interaction and epistasis can be extrapolated to devise a conceptual framework for studying the combined effects of multiple concurrent stimuli. There are several benefits of using this approach; (1) all the genetic, biochemical, and computational tools and concepts developed for studying gene interactions would become available for studying the effects of the environment, (2) it would allow for easier mechanistic interpretation of the responses to complex environmental stimuli, (3) the contributions of an individual stimulus to altering biological processes can be more easily elucidated, and (4) it would provide a unifying framework for studying gene-gene, gene-environment and environment-environment interactions.

In this study, we define an environmental interaction as the interaction between different environmental stimuli that affect the same observable characteristic or trait. Similar to the statistical definition of genetic epistasis, environmental epistasis is an environmental interaction in which the effects of the individual stimuli are not independent of each other [20,21,22]. To test our hypothesis, we used the yeast *S. cerevisiae* and grew cells at standard conditions (glucose, 30°C) and changed growth conditions to either high temperature (37°C, HT stimulus) or the non-fermentable carbon source glycerol (G stimulus), and concurrently with both environmental stimuli (glycerol, 37°C, HT+G stimuli) (S1 Fig). Using precise quantitative proteomics of the *S. cerevisiae* proteome and the changes in protein abundance as the readouts of the interactions, we show that environmental interactions and epistasis play central roles in determining the state of the proteome in response to multiple, concurrent environmental stimuli. We also show that, using the dominance of one stimulus over another, environmental interactions can be used to identify proteins that are important for responding to a dominant stimulus. We validated our approach using an independent publicly available transcriptomic dataset.

## Experimental Procedures

### Strains and Media

All experiments used the diploid *S. cerevisiae* strain BY4743, which has been previously described [23]. Cells were grown using standard techniques [24].

### Growth rate analysis

Cells were grown in 96 well plates in 100 μL cultures (10 μL of starter culture and 90 μL of fresh media) with continuous shaking in a BioTek Synergy 4 Hybrid Microplate Reader for 10 h. Growth rates were assayed in 8 conditions: (1) Synthetic complete medium with glucose (ScD) at 30°C, (2) ScD at 37°C, (3) Synthetic complete medium with glycerol (ScG) at 30°C, (4) ScG at 37°C, (5) Yeast extract, peptone medium with glucose (YPD) at 30°C, (6) YPD at 37°C, (7) Yeast extract, peptone medium with glycerol (YPG) at 30°C, and (8) YPG at 37°C. Absorbance was measured at 660nm at 3 min intervals. Using custom *R* scripts, the doubling times were calculated from the linear regression curve through the log growth phase using the log of the absorbance and time of growth. A two-tailed *t*-test of independence with Bonferroni correction for the 11 comparisons (7 comparisons of the control, YPD at 30°C, to the test conditions, 3 comparisons of the observed concurrent double stimuli effect to the expected sum of individual stimulus effects, and 1 comparison of the observed concurrent three stimuli effect to the expected sum of the effects of the three individual stimulus) was used to calculate the statistical significance of a stimulus effect on the growth rate [25].

### Preparation of yeast protein extracts

Five mL of YPD (1% yeast extract, 2% peptone, 2% glucose) was inoculated with a single yeast colony from a YPD agar plate and grown overnight. Three replicates were grown under each growth condition: YPD at 30°C and 37°C and YPG at 30°C and 37°C. Fifty mL of YPD was inoculated with 50 μL of the overnight culture and grown at 30°C and 37°C. One hundred mL of YPG (1% yeast extract, 2% peptone, and 3% glycerol v/v) was inoculated with overnight cultures and grown at 30°C and 37°C. The cultures were grown with constant shaking at 175 rpm in Innova 44 shaker incubators (New Brunswick Scientific). For all four growth conditions, cells were harvested at mid-log phase as determined by $OD_{600}$ measurements. Cells grown in YPD were harvested after 14 h, while cells grown in YPG were harvested after 24 h. All cultures were centrifuged at 2000 rpm for 5 min at 4°C using a Sorvall HLR6/H600A/HBB6 rotor in Sorvall RC-3B centrifuge and washed with ice cold deionized $H_20$. The cell pellets were resuspended in 1 mL ice cold wash buffer (10 mM Tris pH 8.0, 5 mM beta-mercaptoethanol, 500 mM ammonium chloride, 100 mM magnesium acetate) and lysed at 4°C using glass beads and a Bead Beater (BioSpec, Inc) for 10 min as previously described [26]. The whole cell extracts (WCE) were clarified by centrifugation at 20,000g for 15 min at 4°C, and a 200 μL aliquot of the cleared WCE was stored at -80°C.

### Isobaric tag for relative and absolute quantitation (iTRAQ) labeling

The total protein concentration was determined using a Bradford assay according to the manufacturer's protocol (Sigma Aldrich). For each growth condition, 50 μg of total protein was mixed with 50 ng of bovine serum albumin (Thermo Scientific) as an internal standard. Each protein sample was acetone precipitated and resolublized in 25 μl iTRAQ dissolution buffer (500 mM triethylammonium bicarbonate, 0.1% sodium dodecyl sulfate). The proteins were reduced with tris(2-carboxyethyl)phosphine at 60°C for 60 min and the cysteines were derivatized with methyl methanethiosulfonate at room temperature for 10 min. All samples were

digested with sequencing-grade modified trypsin (1:50; Promega Corporation) overnight at 37°C. Equal fractions of the tryptic digests from the three replicates grown in YPD at 30°C were pooled separately and used as a control for the iTRAQ experiments. Fifty μg of the pooled control and 50 μg of each of the replicates were used for iTRAQ labeling. The iTRAQ labeling reagents were resolublized in 150 μL anhydrous ethanol (Sigma Aldrich). 75 μL of iTRAQ reagent solutions were added to each 50 μg sample, incubated with shaking for 1 h at room temperature on an Eppendorf Thermomixer R, pooled, frozen, lyophilized, resolublized in 200 μL of buffer A (0.1% formic acid), and stored at -80°C.

## Liquid chromatography and mass spectrometry

The iTRAQ-labeled samples were analyzed with MudPIT as previously described [27]. Precursor ions were analyzed in the Orbitrap mass analyzer followed by four CID fragment ion scans in the ion trap and four HCD fragment ion scans (normalized collision energy = 45%) in the Orbitrap.

iTRAQ data analysis: RAW files generated by the MudPIT experiments were searched using the Sequest HT database search engine running under Proteome Discoverer v1.4 (Thermo Scientific) against a forward and reverse yeast protein database (S.cererevisiae_orf_trans_all_SGD.fasta.6718) with appended common contaminant sequences [28,29]. Beta-methylthiolation and iTRAQ modifications were included as constant modifications. Oxidation of methionine and tryptophan, and deamidation of glutamine and asparagine were used as variable modifications. Precursor mass tolerance was set to 3 Da and fragment mass tolerance was set to 0.8 Da. Protein assembly, reporter ion quantitation, and protein fold change calculations were done using ProteoIQ at 5% peptide and protein FDR (Premier Biosoft). Hierarchical clustering analysis was done using Cluster 3.0 [30]. Heatmaps were generated using Java Treeview [31]. Circos plots were generated as described in Krzywinski *et. al.* to visualize the genomic locations of the quantitated proteins [32]. For better visualization, only those regions of the genome that were quantitated in this study are shown. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD002371 [33].

## Environmental interaction analysis

All analysis was performed using *R* scripts to parse the fold change expression data to identify proteins that show specific expression patterns in response to complex environmental stimuli. For each protein, we used linear regression to test for any association of high temperature or glycerol using a model that included main effects for glycerol and temperature and the glycerol by temperature interaction. We used the effect size estimates and ANOVA *p-values* (3 degrees of freedom) calculated by the *lm* function and adjusted the *p-values* for a 5% FDR using the Benjamini-Hochberg procedure for finding differentially expressed proteins [34]. We used the adjusted *p-value* cut-off of 0.05 to determine statistical significance. If the overall adjusted *p-value* was greater than 0.05, we classified the proteins as non-responders. The positive and negative signs of the effect size estimates correspond to upregulation and downregulation, respectively, showing the direction of change. The remaining proteins were further classified into environmental interaction classes based upon the effect size estimate *p-values* and the direction of change. If the *p-value* of an estimate was less than 0.05, the protein was considered differentially expressed in response to that environmental stimulus.

To test if a protein is affected by environmental epistasis, the effect size estimates for the individual high temperature (HT) and glycerol stimuli (G) were summed, the combined standard error calculated as the square root of the sum of the squared standard errors, and a

two-sample *t*-test of independence was used to compare the summed effect size estimate to the effect size estimate for the concurrent high temperature and glycerol stimuli (HT+G). If a *t*-test *p-value* was less than 0.05, the protein was assumed to be affected by environmental epistasis.

### Environmental interaction analysis of transcriptomic dataset

Normalized expression data described in Knijnenburg *et. al.* was downloaded [13]. The transcriptomic data were generated using haploid *S. cerevisiae* (CEN.PK113-7D MATa) cells grown in chemostat cultures [13]. We chose 4 culture conditions similar to our experimental design for further analysis. The culture conditions tested were: 1) with ammonium sulfate as the nitrogen source (n = 5), 2) with methionine as the nitrogen source (n = 3), 3) anaerobic conditions (n = 4), and 4) with methionine as the nitrogen source and anaerobic conditions concurrently (n = 3). Transcriptomic data from the cells grown with ammonium sulfate as the nitrogen source were used as the baseline control. The fold change was calculated by subtracting the average normalized expression data of baseline samples from the individual expression data. Finally, the genes were classified into various types of environmental interactions as described above.

### Co-expression network analysis

Sparse PArtial Correlation Estimation (SPACE) was used to build protein co-expression networks and identify the hub genes [35]. To account for outliers, the data were normalized using probabilistic quotient normalization and scaled using a generalized logarithmic scaling factor [36,37]. The data were scaled and centered to have a standard deviation of 1 and mean of 0 to remove any bias in the correlation analysis [38]. We estimated the partial correlation matrix using the space.dew method implemented in the SPACE *R* package [35]. We selected the default value of the tuning parameter for constructing the initial network [35]. The network was visualized in Cytoscape 3.1.1 [39].

## Results

While cells measure and respond to many environmental stimuli, we chose temperature and carbon source to test our hypothesis. Both stimuli are known to be important factors for survival and have wide-ranging effects on yeast metabolism [1]. We used growth with glucose at 30°C as the control, and high temperature and glycerol as the stimuli. The changing growth conditions were: glucose at 37°C (HT stimulus), glycerol at 30°C (G stimulus), and glycerol at 37°C concurrently (HT+G stimulus). To precisely measure the proteomic responses of the cell, we used isobaric tag for relative and absolute quantitation (iTRAQ) labeling followed by multidimensional protein identification technology (MudPIT)-based mass spectrometry to quantify the steady state proteomes under the four different growth conditions (S1 Fig) [40,41]. A total of 1064 proteins were quantitated in the control and the three test conditions. We filtered the data to focus only on the 466 proteins that were quantitated in all three independent replicates of all of the three test conditions (Fig 1A, S1 Table). Cross-correlation analysis of the filtered data showed high reproducibility among the replicates (S2 Fig). The proteomic changes in the cells grown with the concurrent stimuli were more similar to the changes induced by glycerol compared to high temperature (S2 Fig).

We defined the response to an environmental factor(s) as the $\log_2$-fold change in protein abundance/expression between the control and experimental conditions. For this study, we used "fold change" to denote the $\log_2$ fold change. We built linear regression models for each protein using fold changes to estimate the effect sizes of the stimuli. We used ANOVA for estimating statistical significances since we were comparing multiple stimuli. We interpreted the

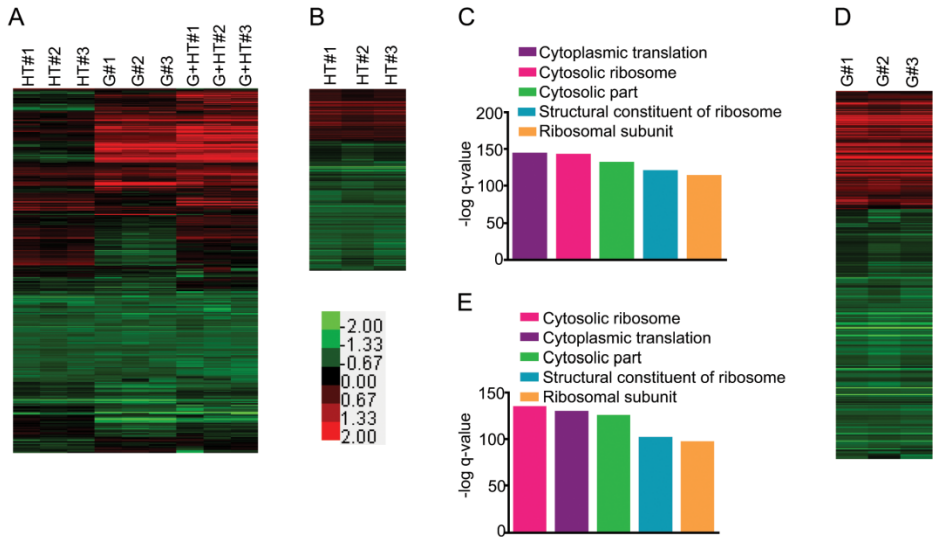



**Fig 1. Proteomic responses to complex environmental stimuli.** Diploid *S. cerevisiae* (BY4743) cells were grown in rich media under 4 conditions: 1) glucose as the carbon source at 30°C, 2) glycerol as the carbon source at 30°C, 3) glucose at 37°C, and 4) glycerol at 37°C. Three biological replicates for each growth conditions were performed. Fold changes were calculated from iTRAQ reporter ion intensities using reporter ion intensities from the pooled replicates of growth in glucose as the carbon source at 30°C as the baseline. The fold changes were $\log_2$ transformed for downstream analysis. The color bar shows the fold change range. A) Complete filtered proteomic dataset for high temperature stimulus (HT), glycerol stimulus (G), and concurrent glycerol and high temperature stimuli (HT+G) (Red: Up, Green: Down, Black: No change). The heatmap represents the fold changes of 466 proteins. B) Fold changes of 283 proteins differentially expressed in response to HT stimulus. C) Bar graph shows the–log q-value of enrichments of the top 5 pathways in the list of proteins differentially expressed after the HT stimulus. D) Fold changes of 379 proteins differentially expressed in response to the G stimulus. E) Bar graph shows the–log q-value of enrichments of the top 5 pathways in the list of proteins differentially expressed after the G stimulus.

doi:10.1371/journal.pone.0134099.g001

positive or negative sign of the effect size as either upregulation or downregulation, respectively. The Benjamini-Hochberg procedure was used to adjust the ANOVA *p-values* at 5% FDR [34]. A protein was assumed to be differentially expressed if the adjusted overall ANOVA *p-value* was less than 0.05. These proteins were further analyzed and classified into different environmental interaction classes using the direction of the change (upregulated or downregulated) and the *p-values* of the effect size estimates [34].

## Stimuli-specific expression patterns can be used to identify proteins important for responding to the stimuli.

We observed 283 proteins differentially expressed with high temperature, 379 proteins differentially expressed in response to glycerol, and 370 proteins were differentially expressed in concurrent high temperature and glycerol (Fig 1B and 1D, and S1 Table), while 41 proteins did not change in response to any of the stimuli. We selected GeneMANIA Cytoscape plugin for pathway analysis since it extends the input list of differentially expressed proteins by adding related proteins to enhance sensitivity and coverage [42,43]. It also allows using the complete proteome as the background. This helped to build a more complete picture of differentially regulated pathways. Pathway analysis of these two differentially expressed protein groups revealed the same top five pathways; none were specific to either stimulus (Fig 1C and 1E). All of the top five pathways were related to protein synthesis and translational control, suggesting that the

regulation of protein synthesis is an important step in responding to environmental stimuli. Translation factors are some of the most abundant proteins in yeast and our proteomic assays are limited by the abundance of proteins in the cell. Although this could have confounded pathway analysis and led to the identification of translation associated pathways as being the most enriched, using only the differentially expressed proteins suggests that these pathways are, at the least, being differentially regulated. Furthermore, similar observations have also been made in previous studies [1,10,44]. It is noteworthy that the pathways expected to be important for responding to these stimuli, such as "protein folding" for growth at high temperature and "TCA cycle" for growth with glycerol were present farther down the list at numbers 39 and 53, respectively (S2 and S3 Tables) [45,46,47]. This mirrors a common problem in 'omics' studies that generate large lists of candidate genes, transcripts and proteins. The important responders are lost in a long list where a majority of differentially expressed genes or proteins is not directly responding to the stimulus. Therefore, choosing candidates for an in-depth mechanistic study becomes a challenge.

To address this problem, we devised a methodology using dominance in environmental interactions to identify proteins and pathways important for responding to a stimulus. We noticed proteomic expression patterns in which the response to one stimulus was dominant over the other. We speculated that a protein critical in responding to a stimulus will respond to that stimulus even when challenged by a competing stimulus. If this hypothesis is correct, such an environmental interaction could serve as a filter to select and identify proteins that respond specifically to the dominant environmental stimulus.

To test this hypothesis, we classified the list of 466 proteins responding to the concurrent glycerol and high temperature stimuli based upon their expression patterns. Two classes of dominant environmental interactions are possible. In one class, a stimulus reverses an expression change induced by the other stimulus (Fig 2A and 2B, top panels, rows 1 and 3). In the other class, a stimulus induces a change in expression, while the other stimulus has no effect on its own and does not change the response to the concurrent stimulus (Fig 2A and 2B top panels, rows 2 and 4). Each class is represented by two theoretical expression patterns for a total of four expression patterns for each stimulus (Fig 2A and 2B top panels).

For the environmental interactions in which the HT stimulus was dominant over the G stimulus, the *p-values* for all of the effect size estimates were less than 0.05. The changes for the HT and HT+G stimuli were in the same direction and differed from the G stimulus (Fig 2A, top panel, rows 1 and 3). Alternatively, the *p-values* for only the HT and HT+G stimuli effect size estimates were less than 0.05 and the directions of change for the HT and HT+G stimuli were the same (Fig 2A, top panel, rows 2 and 4). In all, we identified 30 proteins for which the response to the HT stimulus was dominant over the G stimulus (Fig 2A and S1 Table). We used pathway analysis to identify which protein classes were responding to the dominant stimulus. The group of proteins for which the HT stimulus was dominant included the heat shock response proteins HSP10, HSP60, SSA1, SSA2, and HSP150 (Fig 2A bottom panel, and S1 Table). Pathway analysis of these 30 proteins showed that the top five enriched pathways included protein folding, protein refolding, and unfolded protein binding (Fig 2C, S5 Table). These pathways are expected to be important for growth at higher temperatures [45,46,48,49].

For the environmental interaction in which the G stimulus is dominant, we saw a similar set of patterns as described above except the G stimulus dominates the HT stimulus (Fig 2B, top panel). There are 121 proteins for which the response to the G stimulus was dominant over the HT stimulus (Fig 2B, bottom panel and S1 Table). The group of proteins for which the G stimulus was dominant includes metabolic enzymes such as CDC19, ACO1, and LSC1. (Fig 2B, bottom panel, and S1 Table). Pathway analysis of these 121 proteins showed that the top five pathways included the oxidation-reduction process, the generation of precursor metabolites

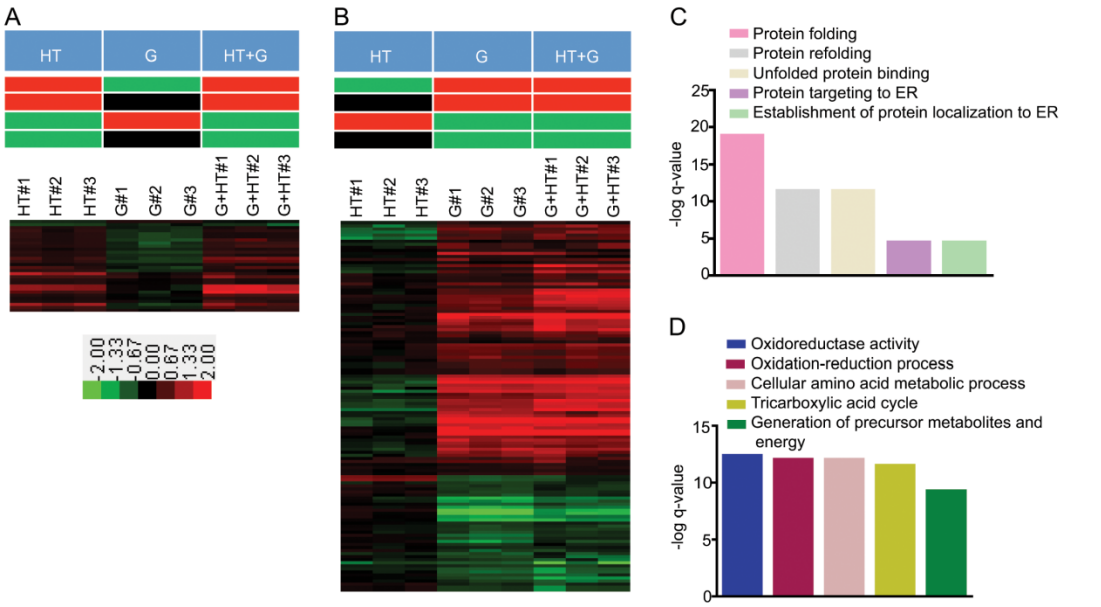



**Fig 2. Dominance of an environmental stimulus used to identify proteins that are important for responding to the environmental stimulus.** The color bar shows the range of fold changes. Pathway analysis was done using the *GeneMANIA* Cytoscape plugin [42]. Bar graphs were generated in *Graphpad Prism*. A) Proteins for which HT stimulus is dominant over G stimulus. The theoretical expression patterns are depicted in the top panel (Red, upregulation; green, downregulation; and black no statistically significant change in expression). The heatmap of fold changes in expression for 30 proteins for which HT stimulus is dominant over G stimulus is shown in bottom panel. B) Proteins for which G stimulus is dominant over HT stimulus. The theoretical expression patterns are depicted in the top panel. The heatmap of fold changes in expressions for 121 proteins for which G stimulus is dominant over HT stimulus is shown in bottom panel. C) Bar graph shows the–log q-value of enrichments of the top five pathways in the list of proteins for which HT stimulus is dominant over G stimulus. D) Bar graph shows the–log q-value of enrichments of the top five pathways in the list of proteins for which G stimulus is dominant over HT stimulus.

doi:10.1371/journal.pone.0134099.g002

and energy, and the tricarboxylic acid cycle (Fig 2D, and S6 Table). All of these three pathways are expected to be important for respiratory growth [47,50,51]. Consistent with our hypothesis, pathway analysis of proteins that respond to a dominant environmental stimulus reveals a functional relationship to the response to the stimulus. High temperature has a dominant effect on proteins involved in protein folding, while glycerol has a dominant effect on proteins involved in respiratory metabolism. These results show the practical applications of using dominant environmental interactions to identify proteins that respond to specific stimuli and that are directly involved in the cell's response to that stimulus.

## Analysis of expression patterns reveals that environmental interactions mirror gene interactions.

In addition to the dominant interactions of concurrent environmental stimuli, we observed other classes of environmental interactions that mirror gene interactions. First, we observed a class of proteins whose abundance either increased or decreased in response to both the individual stimuli as well as the concurrent stimuli (Fig 3A). This is similar to gene pairs in which both the individual mutants as well as the double mutant have the same phenotype. We

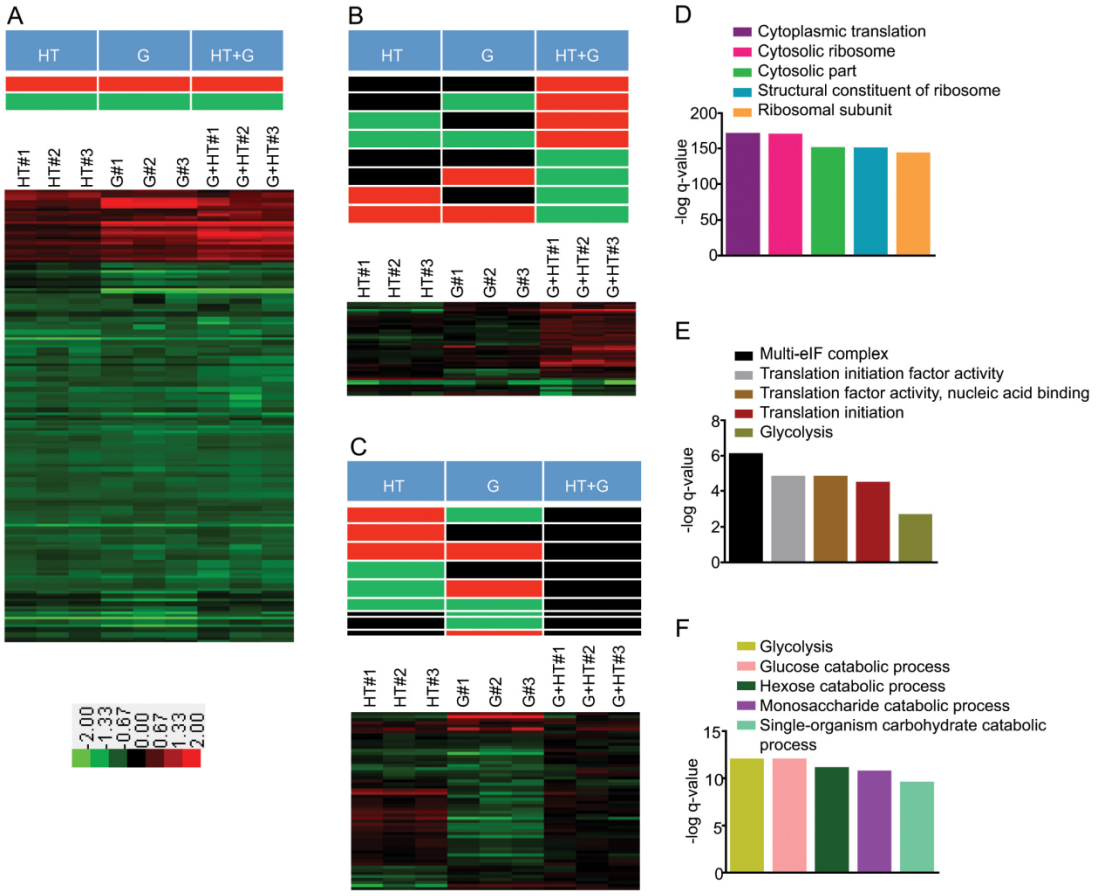



**Fig 3. Proteins in different environmental interaction classes and the corresponding enriched pathways after concurrent G and HT stimuli.** The color bar shows the range of fold changes. Pathway analysis was done using *GeneMANIA* Cytoscape plugin[42]. Bar graphs were generated in *Graphpad Prism*. A) Non-specific environmental response (NER) proteins to individual and concurrent HT and G environmental stimuli. The theoretical expression patterns are shown in the top panel. The fold changes of 175 NER proteins are shown as a heatmap. B) The theoretical expression patterns for discordant environmental interaction are shown in the top panel. The fold changes of 41 proteins are shown as a heatmap. C) The theoretical expression patterns for suppression environmental interaction are shown in the top panel. The fold changes of the 58 proteins affected by suppression are shown as a heatmap. D) Bar graph shows the–log q-value of enrichments for the top 5 pathways for the non-specific environmental response proteins. E) Bar graph shows the–log q-value of enrichments for the top 5 pathways in the list of proteins affected by discordant environmental interaction. F) Bar graph shows the–log q-value of enrichments for the top 5 pathways in the list of proteins affected by suppression environmental interaction.

doi:10.1371/journal.pone.0134099.g003

classified these proteins as non-specific environmental responders. This class is represented by two theoretical expression patterns: activated or repressed (Fig 3A, top panel and S1 Table). For these non-specific environmental response modules, the *p-values* for all the effect size estimates were less than 0.05 and the directions of change were the same (Fig 3A, top panel). We identified 175 proteins that correspond to these patterns, and pathway analysis revealed that

they are largely involved in protein synthesis and translational control (Fig 3A, bottom panel and 3D, and S7 Table).

We also observed proteomic responses to concurrent environmental stimuli similar to gene interactions in which the two single mutants are wild-type or have one phenotype, while the double mutant has a different phenotype (Fig 3B). This class includes proteins whose expression was either decreased or unchanged after a single stimulus but was increased if both stimuli were applied concurrently. The class also includes proteins whose expression was either increased or unchanged after a single stimulus but was decreased by the concurrent stimuli. We classified this environmental interaction group as a discordant class. There are eight theoretical expression profiles in the discordant environmental interaction class (Fig 3B, top panel). For the discordant environmental interaction, the *p-value* for the HT+G concurrent stimuli effect size estimate was less than 0.05 and the directions of change for either the HT or G stimuli were not the same as HT+G. We identified 41 proteins that show discordance (Fig 3B, bottom panel and S1 Table). They are mainly involved in protein synthesis and metabolic pathways (Fig 3E, and S8 Table).

Finally, we observed suppression, in which a protein's abundance changed in response to a single stimulus, yet the change was suppressed by the simultaneous application of the second stimulus (Fig 3C). This class is similar to gene interactions in which double mutants show the wild-type phenotype [52,53]. The suppression class is represented by eight theoretical expression patterns (Fig 3C, top panel). For suppression environmental interactions, the *p-value* for the HT+G effect size estimate was more than 0.05, and the *p-value* for at least one of HT and G stimuli effect size estimates was less than 0.05. We identified 58 proteins that are affected by suppression (Fig 3C, bottom panel and S1 Table). Pathway analysis revealed that metabolic pathways are most affected by suppression (Fig 3F and S9 Table).

## A large fraction of the proteome is affected by environmental epistasis

An important feature of genetic epistasis is that the modulating effects of multiple genes are not always independent of each other [20,21,22,54,55]. In many cases, non-independence is diagnostic of a functional relationship between genes [20,22,54]. Genetic epistasis is used to test if the effects of genetic elements are independent. Genetic epistasis occurs when the effects are not independent. We tested if the effects of these two individual environmental stimuli were independent of each other for individual proteins in the proteome. Similar to the mathematical approach to genetic epistasis, we measured the response of each protein and classified a response as influenced by environmental epistasis if the sum of the effects of the individual stimuli for a protein was not equal to the response to the concurrent stimuli (*t*-test, *p*-value $\leq$0.05) [20,21,22]. We used $\log_2$ fold change as the measure of the effect of a stimulus. From our list of 466 quantitated proteins, 240 proteins were affected by environmental epistasis (S1 Table). Pathway analysis of these proteins revealed that a majority of the enriched pathways are involved in protein synthesis and translational control (Fig 4A and S10 Table). The topmost enriched pathways included cytoplasmic translation, cytosolic ribosome, and structural constituent of ribosome (Fig 4A and S11 Table).

Pathway analysis of the 226 proteins not affected by environmental epistasis revealed a large number of metabolic pathways (Fig 4B and S1 and S10 Tables). It is interesting to note that the distribution of pathways affected by environmental epistasis is different from those that are unaffected. Protein synthesis and translational control seem to be disproportionately affected by environmental epistasis compared to other pathways. These pathways have previously been found to change in response to the changes in the growth rate [56,57]. If the effects of the two stimuli on the growth rate are not independent, it could explain the observed environmental

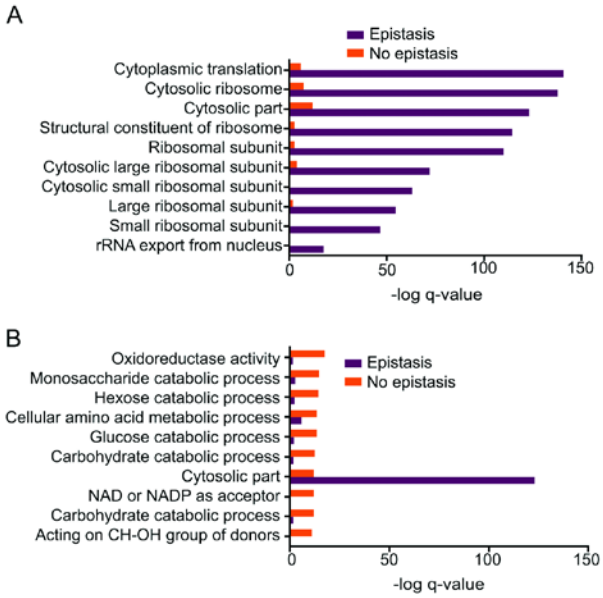



**Fig 4. Environmental epistasis in the proteomic response to concurrent stimuli.** Pathway analysis was done using the *GeneMANIA* Cytoscape plugin[42]. Bar graphs were generated in *Graphpad Prism*. A) Bar graph shows the–log q-value of enrichments of the top 10 pathways in the list of proteins affected by epistasis (purple) and their–log q-value in the list of proteins not affected by epistasis (orange). B) Bar graph shows the–log q-value of enrichments of the top 10 pathways in the list of proteins not affected by epistasis (orange) and their–log q-value in the list of proteins affected by epistasis (purple).

doi:10.1371/journal.pone.0134099.g004

epistasis. To test the independence in the effects of the two stimuli on the growth rate, we determined the doubling times under the same conditions. The change in the doubling times was used to measure the effect of a stimulus. Our data shows that the effects of high temperature and glycerol on the growth rate are additive and, therefore, independent of each other (S4 Fig). Further studies are required to elucidate the functional significance of the environmental epistasis.

A number of genetic epistasis subtypes have been defined based upon the mathematical models used to measure the expectation of a phenotype in double mutants [54,55,58,59,60]. Four most commonly used definitions are (1) additive, (2) multiplicative, (3) minimum, and (4) log [55,58]. Although we used only the additive definition for developing the concept of the environmental epistasis in this study, future studies can be performed to compare the results obtained using different definitions.

## Environmental interactions and epistasis regulate mRNA levels.

Although, we identified the environmental interactions using quantitative proteomic data, we speculated that this framework would be applicable to any quantifiable readout including transcriptomic and phenotypic traits. In pioneering studies using chemostat cultures of *S. cerevisiae*, Knijnenburg *et al.* measured the transcriptional response of yeast to multiple, concurrent environmental stimuli [13]. They found linear regression models of expression for the vast

majority of genes required a combinatorial interaction term[13]. This suggests the change in transcription of most genes cannot be explained by simply adding the effects of the individual stimuli. Based on our proteomic results, we hypothesized that environmental epistasis plays a role in determining the state of the transcriptome as well.

To test if our environmental interaction and epistasis models are observed in the transcriptomic responses to concurrent stimuli, we analyzed Knijnenburg dataset which measured the transcriptomic responses of yeast cells growing in carbon limited chemostat cultures [13]. In the experiment, two concurrent stimuli were applied: (1) a change in nitrogen source from ammonium sulfate to methionine and (2) a change from aerobic to anaerobic growth (Fig 5A and S12 Table) [13]. The data showed 564 transcripts were affected by environmental epistasis, while 5987 transcripts were not affected (*p-value* $\leq$0.05) (S12 Table). In contrast to our proteomic analysis, pathway analysis of the transcripts affected by environmental epistasis revealed enrichment for pathways including microbody, peroxisome, and phytosteroid metabolic process (S13 Table). This could be because of the differences between the strains, stimuli, and culture conditions used in the transcriptomic and our proteomic studies. Similar to our proteomic analysis, we observed dominant environmental interactions in the expression of the transcripts (Fig 5B and 5C and S12 Table). Nitrogen source was dominant for 281 transcripts (Fig 5B and S12 Table). Pathway analysis of these transcripts identified pathways involved in methionine metabolism such as sulfur amino acid metabolic process, sulfur compound metabolic process and methionine metabolic process (Fig 5D and S14 Table). Similarly, anaerobic growth was dominant for 938 transcripts (Fig 5C and S12 Table). Pathway analysis of these differentially expressed transcripts showed enrichment of pathways involved in energy production such as cellular respiration, mitochondrial membrane and respiratory chain (Fig 5E and S15 Table). We also observed the same environmental interaction classes in their transcriptomic data as in our proteomic data, including non-specific environmental response, discordance, and suppression (S12 Table). These results strongly suggest that environmental interactions play a significant role in regulating the biochemical state of cells.

## Coexpression network analysis shows community structures are guided by environmental interaction and epistasis.

Coexpression networks link together proteins whose expression levels are regulated in the same way [61,62]. As a consequence, coexpression network analysis can be used to determine if the abundances of proteins affected by environmental epistasis are regulated differently than the proteins that are not affected by environmental epistasis. To explore the protein modules whose expression changes are correlated with each other, we built a coexpression network using the merged proteomic responses from both individual and concurrent stimuli using the Sparse PArtial Correlation Estimation approach (SPACE) (Fig 1A) [35]. An edge, representing coexpression, was introduced between two proteins if the correlation between them was above the average of the correlation matrix. To validate the network, we first tested the power law structure of the reconstructed network [35,63]. The reconstructed network followed the power law distribution. The power law parameter $\alpha$ was approximately 4, which is close to the empirically observed value of 3.45 [63]. Next, we repeatedly reconstructed the network by varying the tuning parameter around the default value and fitting the network to the power law distribution. We found that the reconstructed network follows the power law distribution and that the power law parameter was in the range of 3.75. We identified the sub-graph spanned by the top 1% of highly connected nodes. We found that the Jaccard similarity score of these highly connected nodes was 0.83 on the scale of 0 to 1. Therefore, these nodes were classified as hub nodes, which is one of the characteristic features of power law networks. There were 7 hub

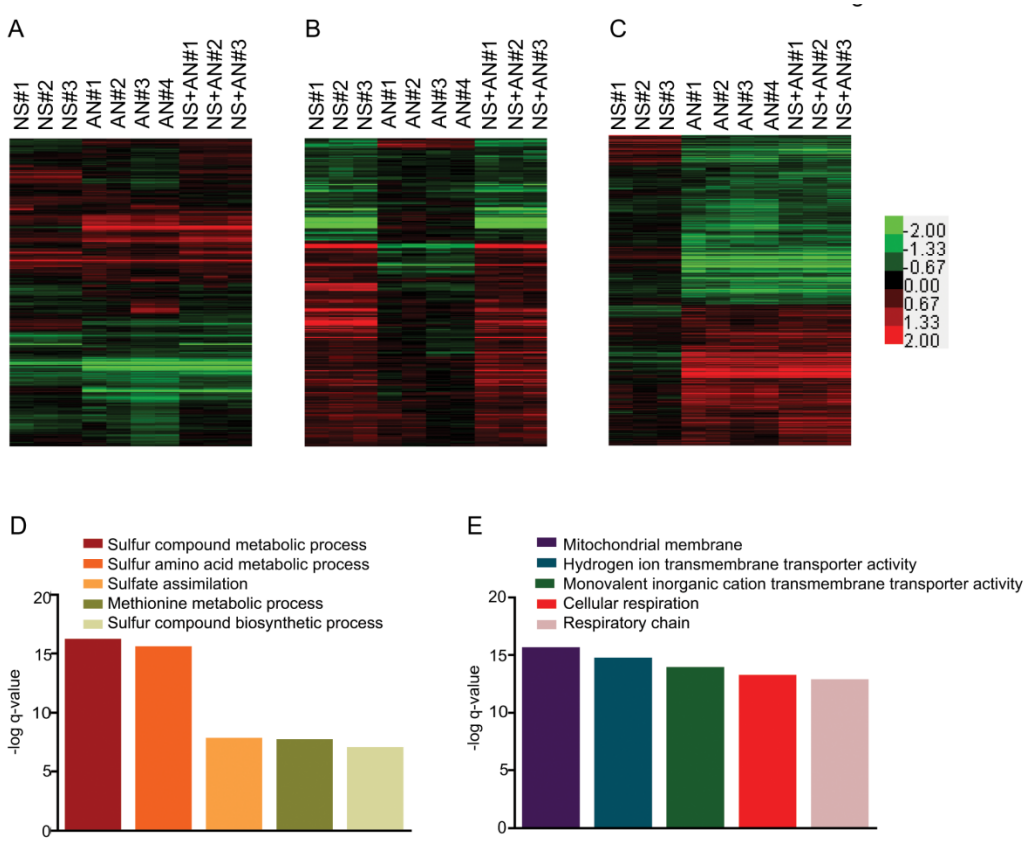



**Fig 5. Environmental interactions affect transcriptomic profiles as well.** Normalized expression data from *Knijnenburg et. al. 2009* was used for the analyses. The transcriptomic data used in the study used haploid *S. cerevisiae* cells (CEN.PK113-7D MATa) grown in carbon limited chemostat cultures under 4 conditions – 1) ammonium sulfate as nitrogen source (n = 5), 2) methionine as nitrogen source, NS stimulus (n = 3), 3) Anaerobic condition, AN stimulus (n = 4), and 4) methionine as nitrogen source under anaerobic conditions NS+AN stimulus (n = 3) [13]. Fold changes were calculated from normalized expression data using average normalized expression data from the five replicates of growth with ammonium sulfate as the baseline. The color bar shows the range of fold changes. Pathway analysis was done using *GeneMANIA* Cytoscape plugin[42]. Bar graphs were generated in *Graphpad Prism*. A) A heatmap of fold changes of the complete transcriptomics dataset consisting of 6551 transcripts. B) A heatmap showing the fold changes for 281 transcripts for which NS stimulus is dominant. C) The–log q-value of enrichment for the top 5 pathways enriched in the list of transcripts for which NS stimulus is dominant. As anticipated, pathways expected to be involved in metabolization of methionine are enriched. D) A heatmap showing the fold changes for 938 transcripts for which AN stimulus is dominant. E) The–log q-value of enrichment for the top 5 pathways enriched in the list of transcripts for which AN stimulus is dominant. As anticipated, pathways expected to be involved in energy production are enriched.

doi:10.1371/journal.pone.0134099.g005

nodes based upon the above criterion. Next, we checked the significance of the identified hubs using the Wilcox Rank sum test and found that the hub community is statistically significant ($p$-value = 0.04) [64]. Finally, we compared the reconstructed network with BioGrid protein interaction data and found that approximately 30% of the edges are previously known interactions and that these interactions were found in every reconstructed network when we varied

the tuning parameter to estimate the partial correlation matrix [65]. The final coexpression network consisted of 329 nodes with at least one neighbor and a total of 359 edges (Fig 6A).

The largest community within this network includes 205 nodes and 249 edges, with two clearly separate sub-graphs connected by a single node (Fig 6B). Interestingly, one sub-graph consists predominantly of proteins affected by environmental epistasis while the second sub-graph consists of proteins not affected by environmental epistasis. Within the global coexpression network, we observed that proteins affected by epistasis were more likely to be linked with each other than with proteins that are not affected by epistasis and vice versa (Fig 6A). There are 199 edges between two proteins affected by epistasis and 85 edges between two proteins not affected by epistasis. However, only 75 edges involved proteins of both types (Fig 6C). This structural organization of the coexpression network suggests that the responses of proteins affected by environmental epistasis are controlled by a different mechanism than the responses of those not affected by environmental epistasis.

Previous studies indicate that proteins linked in a coexpression network are likely to function in the same pathway [61]. We hypothesized that the grouping of proteins upon classification into environmental interaction classes might be driven by their functional associations. If true, we would expect to find more edges in the coexpression network between proteins within the environmental classes. Indeed, we found this result in this network. Our data show that 299 of the edges (83%) are between proteins in the same environmental interaction class, while only 60 are between proteins in different classes (Fig 6D).

## Discussion

Using the concepts of gene interactions and epistasis, we have developed a unifying conceptual framework to understand the cellular responses to complex environmental stimuli. Although, we have only explored the cases with complete dominance of a stimulus, it is possible that both the stimuli contribute to a change in expression. It is also possible that many stimuli contribute towards a change. We speculate that the tools and approaches developed for gene-gene interactions involving multiple genes can be applied in such cases [66]. In addition to linear regression modeling and ANOVA, we also tested our hypothesis using one sample and two sample *t*-tests of independence (data not shown). The results from both approaches were in good agreement.

The effect of mixtures of compounds has been actively studied in toxicology, especially in the context of environmental toxins [67,68,69,70,71,72,73,74,75,76,77,78]. These studies have led to the development of three complementary models to predict the combined effects of compounds in a mixture: (1) in the concentration addition model the total toxicity of a mixture is the sum of the individual toxicities of the component compounds, (2) in the independent action model the toxicities of the components of a mixture are independent of each other, and (3) in the simple interaction model the individual components, at the concentrations being tested, are not toxic, but are toxic when used together in a mixture. These models have been successful in predicting the total toxic effects of mixtures of compounds in many cases [67,69,70,71,72,74,78]. However, it is not immediately clear which one to apply in a specific case without model fitting [69].

Environmental interactions and epistasis can be extrapolated to explain the three models. For example, the concentration addition model can be the case of incomplete dominance where many stimuli affect the biological processes under investigation. This would happen if the compounds in the mixture affect similar biological pathways. If the actions of the compounds are antagonistic to each other, it may lead to either the dominance or the suppression interaction. If their actions are not antagonistic, the combined effect would be the sum of the individual effects which could be observed as the non-specific environmental response.

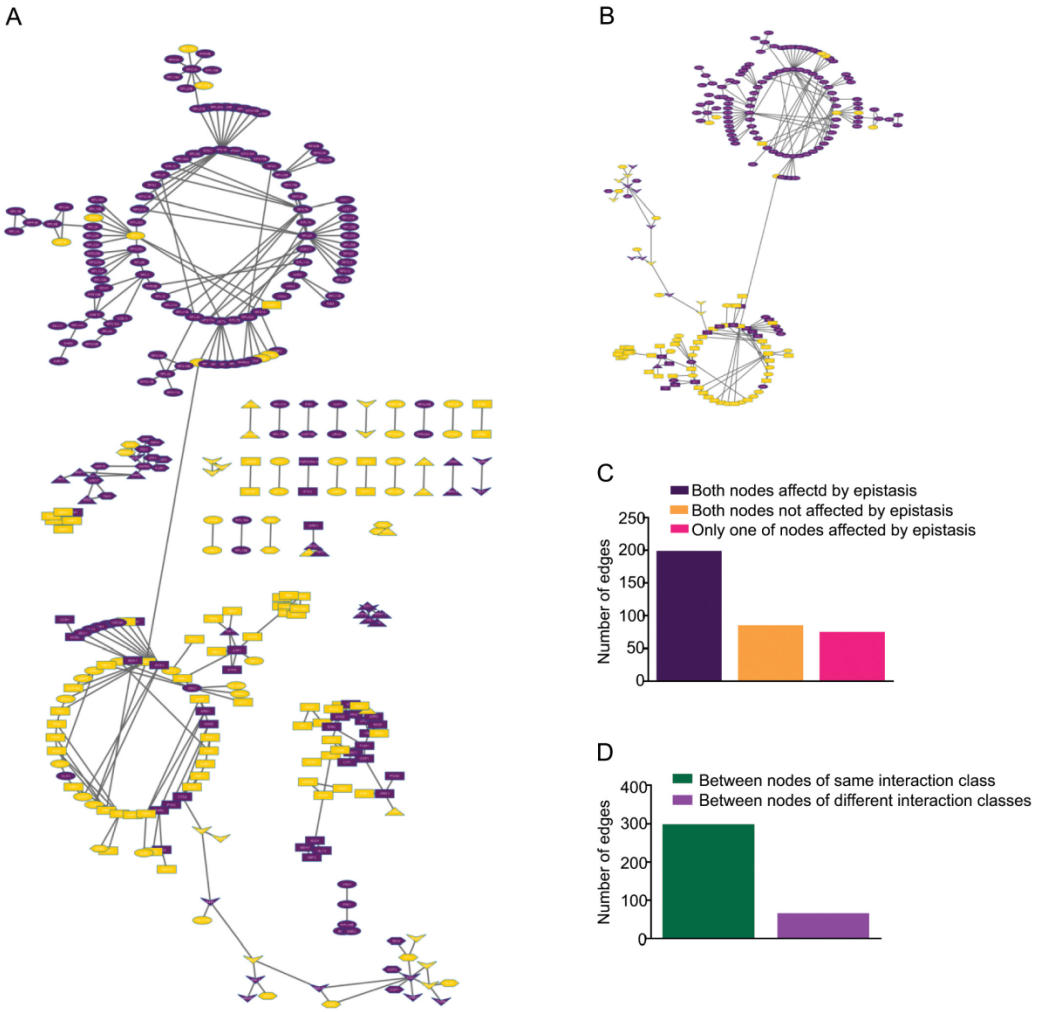



**Fig 6. Coexpression network based on all the quantified proteins and all conditions (G stimulus, HT stimulus, and HT+G stimulus).** Proteins are depicted as nodes. Nodes that are coexpressed are connected with an edge. The coexpression network was generated with *SPACE* algorithm using fold change [35]. Network visualization and analysis was done in Cytoscape 3.1.1 [39]. Bar graphs were generated in *Graphpad prism*. A) All nodes that have at least one edge. Nodes affected by environmental epistasis are highlighted in purple. The circular layout was used to generate the initial network graphics in Cytoscape 3.1.1. Far-flung communities of inter-connected nodes were manually brought together, while preserving the inner community structure, for better visualization. B) The largest community in the coexpression network. Most of the proteins affected by environmental epistasis are members of a subgraph (top circle) that predominantly contains other proteins that are also affected by environmental epistasis. A similar trend is observed with the proteins not affected by environmental epistasis (bottom circle). C) The numbers of three types edges: 1) both nodes are affected by environmental epistasis (199 edges), 2) neither of the nodes are affected by environmental epistasis (85 edges), and 3) only one of the nodes is affected by environmental epistasis (75 edges). Proteins affected by epistasis are predominantly connected to proteins that are also affected by epistasis. D) Number of edges that connect nodes to other nodes within the same environmental interaction classes (299) or between the classes (60). Co-regulatory connections between proteins are predominantly between those of the same class.

doi:10.1371/journal.pone.0134099.g006

The independent action model explains the case where the compounds under investigation act upon different pathways [68,70,71,72,75,77]. This is similar to a gene interaction where two mutations have two unrelated phenotypes and both phenotypes persist in the double mutant. By applying the logic of environmental interaction to this model, we can deduce that the changes induced by a mixture that follows the independent action model would have elements specific to the component compounds of the mixture. Additionally, the changes important to a specific compound would persist in the combinatorial condition, which could be used to identify molecules and pathways that respond to the specific compound in the mixture.

The simple interaction model explains the cases where the compounds individually have little or no toxicity, but are toxic when applied together [73,75]. In terms of environmental interaction, this could be a case of the discordance interaction. The effects explained by this model could also be a special case of environmental epistasis, where the combined effect of compounds is more than the sum of their individual effects. It is worth noting that although we discuss only three of the mixture toxicity models, there are a number of other models that explain the toxicities of compounds in a mixture [67,68,69,70,71,72,73,74,75,76,77,78]. Environmental interactions and epistasis provides a conceptual framework unifying the different toxicity models. The interpretation of results can be made simpler using environmental interactions and epistasis.

Phenotypic plasticity provides the conceptual framework for studying the interaction between genotype and environment. Phenotypic plasticity is the ability of an organism to change its phenotype in response to changes in the environment [79]. It has been used to explain the ability of the same genotype to generate different phenotypes in different environments [79]. However, phenotypic plasticity considers the environment as a monolithic entity. It fails to separate the relative contributions of the different environment components, for example; physical components such as temperature and pressure, chemical components such as nutrients, and signaling molecules that activate different pathways. Applying environmental interactions and epistasis would help parse out the individual contributions of the stimuli towards the change in the phenotype.

Similar to genetic epistasis, our data show that the effects of individual environmental stimuli are not necessarily additive. Proteins affected by environmental epistasis are distributed throughout the genome and do not appear to be clustered at specific locations in the genome (S3 Fig). The prevalence of environmental epistasis in determining the changes in the proteome suggests that epistasis needs to be taken into account when building mathematical models of gene expression.

Consideration of environmental epistasis is especially important in light of the recent attempts to build quantitative linear regression models of gene expression in which the independent variables are the environmental stimuli and the dependent variable is gene expression [80]. Interestingly, in a linear regression modeling study of transcriptional regulation in rice under native conditions, the regression model was able to predict gene expression under native conditions even if the environmental parameters varied slightly from those used for building the model. However, the predictive power of the regression model was reduced under controlled laboratory conditions suggesting that there may have been unknown epistatic interactions in the native conditions absent in the controlled lab conditions [80].

Concurrently applied environmental stimuli behave similarly to genetic elements in the way they interact to regulate the biochemical states of the cells. The observation of environmental interactions and epistasis in determining the states of both the proteome and transcriptome in diverse experimental conditions suggests the prevalence of this phenomenon in nature. Essentially, environmental interaction in concert with phenotypic plasticity and gene interactions can be envisaged as a mathematical operator with three components that determines the

changes in the biochemical state of the cell. The gene interaction component is derived from the effects of the genetic elements, while the environmental interaction component results from the effects of all the environmental stimuli. When the gene and environmental interactions are not independent of each other, phenotypic plasticity accounts for the deviations of the observed from the expected characteristic or trait. Most studies so far have treated phenotypic plasticity, gene interactions, and environmental interactions separately due to a lack of a common unifying framework [20,22,53,54,55,68,69,70,72,75,79,81,82,83,84,85]. Our data suggest that as an abstraction, environmental stimuli can be treated as genes to build a conceptual framework that combines the effects of genes and stimuli. Environmental interactions and epistasis play a critical role in cellular homeostasis as seen in this study's patterns of change in the proteome and the transcriptome.

Our data also suggest that a protein or a transcript is more likely to be critical for responding to a dominant environmental stimulus than to a recessive one. This could lead to more efficient experiment designs for identifying factors directly affected by an environmental stimulus. For example, experiments could be designed in which an unrelated stimulus B is applied concurrently with the stimulus of interest A. The proteins or transcripts, for which the effect of A is dominant, would be more likely to be directly affected by stimulus A. We speculate that the same approach may be extended to genetic perturbations. In this case, an environmental stimulus could be applied in conjunction with the genetic perturbation. As with two concurrent environmental stimuli, a transcript or a protein for which the genetic perturbation is dominant may be more likely to be directly affected by it. Therefore, using dominance, environmental interactions can also be used to devise studies to identify agents, such as regulatory RNAs, proteins, or small molecules which are critical for driving a range of biological processes in health and disease including drug interactions, adaptation in tumor microenvironment and immune responses.

## Supporting Information

**S1 Fig. Experimental design workflow used in this study.** Two environmental stimuli used were high temperature and glycerol as the carbon source. Diploid *S. cerevisiae* cells (BY4743) were grown in rich media under 4 conditions: 1) glucose at 30°C (used as control), 2) glycerol at 30°C (G stimulus), 3) glucose at 37°C (HT stimulus), and 4) glycerol at 37°C (HT+G stimuli). Three biological replicates were performed for each condition.
(EPS)

**S2 Fig. Correlation matrix heatmap (red is high).** Correlation matrix was generated in *R*. There is a high correlation among replicates showing reproducibility across experimental replicates.
(EPS)

**S3 Fig. Visualization of *S. cerevisiae* genomic locations of the proteins quantitated with fold changes represented as a heatmap using Circos plot (Red: Up, Green: Down, Black: No change)[32].** Outermost circle- chromosomes, Second circle-fold changes of proteins with HT stimulus, Third circle-fold changes of proteins with G stimulus, Fourth circle-fold changes of proteins with HT+G stimuli, innermost circle-whether affected by epistasis or not (Purple: Affected by environmental epistasis, Orange: Not affected by environmental epistasis).
(EPS)

**S4 Fig. The effect of high temperature and glycerol on yeast doubling times.** Doubling times were calculated for growth in control (n = 25), high temperature (n = 25), glycerol (n = 25), and concurrent high temperature and glycerol (n = 24). The difference in doubling times from

the control was used to measure the effect of the stimuli and is plotted on Y-axis. HT leads to a decrease of -11 minutes (sd = 6), G leads to an increase of 137 minutes (sd = 14), and HT+G leads to an increase of 142. minutes (sd = 22). The expected effect of HT+G was calculated by summing the observed effects of HT and G (Sum HT+G, increase of 127 minutes with sd of 16). The difference in the means for HT+G and Sum HT+G was not statistically significant (*p-value* = 0.1034, two-tailed *t*-test of independence with Bonferroni correction for 11 comparisons)
(EPS)

**S1 Table. Complete data matrix of proteins.**
(TXT)

**S2 Table. GeneMANIA pathway analysis output for HT stimulus.**
(XLSX)

**S3 Table. GeneMANIA pathway analysis output for G stimulus.**
(XLSX)

**S4 Table. GeneMANIA pathway analysis output for HT+G stimulus.**
(XLSX)

**S5 Table. GeneMANIA pathway analysis output for HT stimulus dominance.**
(XLSX)

**S6 Table. GeneMANIA pathway analysis output for G stimulus dominance.**
(XLSX)

**S7 Table. GeneMANIA pathway analysis output for non-specific environmental reponse in protein expression.**
(XLSX)

**S8 Table. GeneMANIA pathway analysis output for discordance in protein expression.**
(XLSX)

**S9 Table. GeneMANIA pathway analysis output for suppression in protein expression.**
(XLSX)

**S10 Table. GeneMANIA pathway analysis output for environmental epistasis in protein expression.**
(XLSX)

**S11 Table. GeneMANIA pathway analysis output for no environmental epistasis in protein expression.**
(XLSX)

**S12 Table. Complete data matrix of transcripts**
(TXT)

**S13 Table. GeneMANIA pathway analysis output for environmental epistasis in transcript expression.**
(XLSX)

**S14 Table. GeneMANIA pathway analysis output for dominance of NS**
(XLSX)

**S15 Table. GeneMANIA pathway analysis output for dominance of AN**
(XLSX)

**S16 Table. Doubling times under the 8 growth conditions.**
(XLSX)

## Acknowledgments

AJL and PS were supported by NIH grant GM064779 and Vanderbilt University School of Medicine IDEAS Program grant 1-04-066-9530 to AJL. We thank Elizabeth M. Link and Allison C. Galassie for their thoughtful comments in the preparation of this manuscript. We thank the Borden Lacy lab for the use of the BioTek Synergy 4 Hybrid Microplate Reader.

## Author Contributions

Conceived and designed the experiments: PS AJL. Performed the experiments: PS. Analyzed the data: PS R JCS AJL. Contributed reagents/materials/analysis tools: PS R JCS. Wrote the paper: PS R JCS AJL.

## References

1. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241–4257. PMID: 11102521
2. Gerner C, Vejda S, Gelbmann D, Bayer E, Gotzmann J, et al. (2002) Concomitant determination of absolute values of cellular protein amounts, synthesis rates, and turnover rates by quantitative proteome profiling. Mol Cell Proteomics 1: 528–537. PMID: 12239281
3. Pratt JM, Petty J, Riba-Garcia I, Robertson DH, Gaskell SJ, et al. (2002) Dynamics of protein turnover, a missing dimension in proteomics. Mol Cell Proteomics 1: 579–591. PMID: 12376573
4. Soufi B, Kelstrup CD, Stoehr G, Frohlich F, Walther TC, et al. (2009) Global analysis of the yeast osmotic stress response by quantitative proteomics. Mol Biosyst 5: 1337–1346. doi: 10.1039/b902256b PMID: 19823750
5. Yan SP, Zhang QY, Tang ZC, Su WA, Sun WN (2006) Comparative proteomic analysis provides new insights into chilling stress responses in rice. Mol Cell Proteomics 5: 484–496. PMID: 16316980
6. Hazlehurst LA, Landowski TH, Dalton WS (2003) Role of the tumor microenvironment in mediating de novo resistance to drugs and physiological mediators of cell death. Oncogene 22: 7396–7402. PMID: 14576847
7. Trédan O, Galmarini CM, Patel K, Tannock IF (2007) Drug Resistance and the Solid Tumor Microenvironment. Journal of the National Cancer Institute 99: 1441–1454. PMID: 17895480
8. Vaupel P, Kallinowski F, Okunieff P (1989) Blood Flow, Oxygen and Nutrient Supply, and Metabolic Microenvironment of Human Tumors: A Review. Cancer Research 49: 6449–6465. PMID: 2684393
9. Whiteside TL (2008) The tumor microenvironment and its role in promoting tumor growth. Oncogene 27: 5904–5912. doi: 10.1038/onc.2008.271 PMID: 18836471
10. Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, et al. (2008) Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. Mol Biol Cell 19: 352–367. PMID: 17959824
11. De Nicola R, Hazelwood LA, De Hulster EAF, Walsh MC, Knijnenburg TA, et al. (2007) Physiological and Transcriptional Responses of Saccharomyces cerevisiae to Zinc Limitation in Chemostat Cultures. Appl Environ Microbiol 73: 7680–7692. PMID: 17933919
12. Kanani H, Dutta B, Klapa MI (2010) Individual vs. combinatorial effect of elevated CO2 conditions and salinity stress on Arabidopsis thaliana liquid cultures: comparing the early molecular response using time-series transcriptomic and metabolomic analyses. BMC Syst Biol 4: 177. doi: 10.1186/1752-0509-4-177 PMID: 21190570
13. Knijnenburg TA, Daran JM, van den Broek MA, Daran-Lapujade PA, de Winde JH, et al. (2009) Combinatorial effects of environmental parameters on transcriptional regulation in Saccharomyces cerevisiae: a quantitative analysis of a compendium of chemostat-based transcriptome data. BMC Genomics 10: 53. doi: 10.1186/1471-2164-10-53 PMID: 19173729
14. Knijnenburg TA, de Winde JH, Daran JM, Daran-Lapujade P, Pronk JT, et al. (2007) Exploiting combinatorial cultivation conditions to infer transcriptional regulation. BMC Genomics 8: 25. PMID: 17241460

15. Murray JI, Whitfield ML, Trinklein ND, Myers RM, Brown PO, et al. (2004) Diverse and specific gene expression responses to stresses in cultured human cells. Mol Biol Cell 15: 2361–2374. PMID: 15004229
16. Tai SL, Boer VM, Daran-Lapujade P, Walsh MC, de Winde JH, et al. (2005) Two-dimensional Transcriptome Analysis in Chemostat Cultures: Combinatorial effects of oxygen availability and macronutrient limitation in *Saccharomyces cerevisiae*. Journal of Biological Chemistry 280: 437–447. PMID: 15496405
17. Vaga S, Bernardo-Faura M, Cokelaer T, Maiolica A, Barnes CA, et al. (2014) Phosphoproteomic analyses reveal novel cross-modulation mechanisms between two signaling pathways in yeast. Mol Syst Biol 10: 767. doi: 10.15252/msb.20145112 PMID: 25492886
18. Pierce B (2005) Genetics: A conceptual approach.: W. H. Freeman and Company.
19. Bateson W (1909) Mendel's Principles of Heredity, by Bateson W.: Cambridge [Eng.] University Press,1909.
20. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11: 2463–2468. PMID: 12351582
21. Fisher RA, Sir. (1958) The genetical theory of natural selection. New York: Dover Publications.
22. Phillips PC (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9: 855–867. doi: 10.1038/nrg2452 PMID: 18852697
23. Baker Brachmann C, Davies A, Cost GJ, Caputo E, Li J, et al. (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast 14: 115–132. PMID: 9483801
24. Amberg DC, Burke DJ, Strathern JN (2005) Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual, 2005 Edition. Long Island, New York.: Cold Spring Harbor Laboratory Press.
25. Dunn OJ (1961) Multiple Comparisons among Means. J Am Stat Assoc 56: 52–64.
26. Browne CM, Samir P, Fites JS, Villarreal SA, Link AJ (2013) The Yeast Eukaryotic Translation Initiation Factor 2B Translation Initiation Complex Interacts with the Fatty Acid Synthesis Enzyme YBR159W and Endoplasmic Reticulum Membranes. Mol Cell Biol 33: 1041–1056. doi: 10.1128/MCB.00811-12 PMID: 23263984
27. Hoek KL, Samir P, Howard LM, Niu X, Prasad N, et al. (2015) A cell-based systems biology assessment of human blood to monitor immune responses after influenza vaccination. PLoS One 10: e0118528. doi: 10.1371/journal.pone.0118528 PMID: 25706537
28. Eng JK, Fischer B, Grossmann J, MacCoss MJ (2008) A Fast SEQUEST Cross Correlation Algorithm. Journal of Proteome Research 7: 4598–4602. doi: 10.1021/pr800420s PMID: 18774840
29. Eng JK, McCormack AL, Yates JR (1994) An Approach to Correlate Tandem Mass-Spectral Data of Peptides With Amino-Acid-Sequences in a Protein Database. Journal of the American Society for Mass Spectrometry 5: 976–989. doi: 10.1016/1044-0305(94)80016-2 PMID: 24226387
30. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster Analysis and Display of Genome-Wide Expression Patterns. Proceedings of the National Academy of Sciences 95: 14863–14868.
31. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. Bioinformatics 20: 3246–3248. PMID: 15180930
32. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639–1645. doi: 10.1101/gr.092759.109 PMID: 19541911
33. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, et al. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 32: 223–226. doi: 10.1038/nbt.2839 PMID: 24727771
34. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.
35. Peng J, Wang P, Zhou N, Zhu J (2009) Partial Correlation Estimation by Joint Sparse Regression Models. J Am Stat Assoc 104: 735–746. PMID: 19881892
36. Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem 78: 4281–4290. PMID: 16808434
37. Durbin BP, Hardin JS, Hawkins DM, Rocke DM (2002) A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 18 Suppl 1: S105–110. PMID: 12169537

38. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7: 142. PMID: 16762068

39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504. PMID: 14597658

40. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, et al. (1999) Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol 17: 676–682. PMID: 10404161

41. Ross PL, Huang YLN, Marchese JN, Williamson B, Parker K, et al. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Molecular & Cellular Proteomics 3: 1154–1169.

42. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, et al. (2010) GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26: 2927–2928. doi: 10.1093/bioinformatics/btq562 PMID: 20926419

43. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol 9 Suppl 1: S4. doi: 10.1186/gb-2008-9-s1-s4 PMID: 18613948

44. Roberts G, Hudson A (2006) Transcriptome profiling of Saccharomyces cerevisiae during a transition from fermentative to glycerol-based respiratory growth reveals extensive metabolic and structural remodeling. Molecular Genetics and Genomics 276: 170–186. PMID: 16741729

45. Richter K, Haslbeck M, Buchner J (2010) The Heat Shock Response: Life on the Verge of Death. Molecular Cell 40: 253–266. doi: 10.1016/j.molcel.2010.10.006 PMID: 20965420

46. Riezman H (2004) Why Do Cells Require Heat Shock Proteins to Survive Heat Stress? Cell Cycle 3: 60–62.

47. Schüller H-J (2003) Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. Current Genetics 43: 139–160. PMID: 12715202

48. Åkerfelt M, Morimoto RI, Sistonen L (2010) Heat Shock Factors: Integrators of Cell Stress, Development and Lifespan. Nat Rev Mol Cell Biol 11: 545–555. doi: 10.1038/nrm2938 PMID: 20628411

49. de Nadal E, Ammerer G, Posas F (2011) Controlling gene expression in response to stress. Nat Rev Genet 12: 833–845. doi: 10.1038/nrg3055 PMID: 22048664

50. Brisson D, Vohl M-C, St-Pierre J, Hudson TJ, Gaudet D (2001) Glycerol: a neglected variable in metabolic processes? BioEssays 23: 534–542. PMID: 11385633

51. Nevoigt E, Stahl U (1997) Osmoregulation and glycerol metabolism in the yeast Saccharomyces cerevisiae. Fems Microbiology Reviews 21: 231–241. PMID: 9451815

52. Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C (2009) Systematic mapping of genetic interaction networks. Annu Rev Genet 43: 601–625. doi: 10.1146/annurev.genet.39.073003.114751 PMID: 19712041

53. St Johnston D (2002) The art and design of genetic screens: Drosophila melanogaster. Nat Rev Genet 3: 176–188. PMID: 11972155

54. de Visser JA, Cooper TF, Elena SF (2011) The causes of epistasis. Proc Biol Sci 278: 3617–3624. doi: 10.1098/rspb.2011.1537 PMID: 21976687

55. Mani R, St Onge RP, Hartman JLt, Giaever G, Roth FP (2008) Defining genetic interaction. Proc Natl Acad Sci U S A 105: 3461–3466. doi: 10.1073/pnas.0712255105 PMID: 18305163

56. Regenberg B, Grotkjaer T, Winther O, Fausboll A, Akesson M, et al. (2006) Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae. Genome Biol 7: R107. PMID: 17105650

57. Slavov N, Botstein D (2011) Coupling among growth rate response, metabolic cycle, and cell division cycle in yeast. Mol Biol Cell 22: 1997–2009. doi: 10.1091/mbc.E11-02-0132 PMID: 21525243

58. Gao H, Granka JM, Feldman MW (2010) On the classification of epistatic interactions. Genetics 184: 827–837. doi: 10.1534/genetics.109.111120 PMID: 20026678

59. Hallgrimsdottir IB, Yuster DS (2008) A complete classification of epistatic two-locus models. BMC Genet 9: 17. doi: 10.1186/1471-2156-9-17 PMID: 18284682

60. Li W, Reich J (2000) A complete enumeration and classification of two-locus disease models. Hum Hered 50: 334–349. PMID: 10899752

61. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249–255. PMID: 12934013

62. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4: Article17.

63. Clauset A, Shalizi C, Newman M (2009) Power-Law Distributions in Empirical Data. SIAM Review 51: 661–703.
64. Kolaczyk ED, Csárdi G (2014) Statistical Analysis of Network Data with R: Springer.
65. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535–539. PMID: 16381927
66. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nature Reviews Genetics 10: 392–404. doi: 10.1038/nrg2579 PMID: 19434077
67. Hermens J, Leeuwangh P, Musch A (1985) Joint toxicity of mixtures of groups of organic aquatic pollutants to the guppy (Poecilia reticulata). Ecotoxicology and Environmental Safety 9: 321–326. PMID: 4006831
68. Altenburger R, Backhaus T, Boedeker W, Faust M, Scholze M (2013) Simplifying complexity: Mixture toxicity assessment in the last 20 years. Environ Toxicol Chem 32: 1685–1687. doi: 10.1002/etc.2294 PMID: 23843317
69. Belden JB, Gilliom RJ, Lydy MJ (2007) How well can we predict the toxicity of pesticide mixtures to aquatic life? Integr Environ Assess Manag 3: 364–372. PMID: 17695109
70. Altenburger R, Nendza M, Schuurmann G (2003) Mixture toxicity and its modeling by quantitative structure-activity relationships. Environ Toxicol Chem 22: 1900–1915. PMID: 12924589
71. Altenburger R, Scholz S, Schmitt-Jansen M, Busch W, Escher BI (2012) Mixture toxicity revisited from a toxicogenomic perspective. Environ Sci Technol 46: 2508–2522. doi: 10.1021/es2038036 PMID: 22283441
72. Altenburger R, Walter H, Grote M (2004) What contributes to the combined effect of a complex mixture? Environ Sci Technol 38: 6353–6362. PMID: 15597892
73. Berenbaum MC (1989) What is synergy? Pharmacol Rev 41: 93–141. PMID: 2692037
74. Deneer JW (2000) Toxicity of mixtures of pesticides in aquatic systems. Pest Management Science 56: 516–520.
75. Greco WR, Bravo G, Parsons JC (1995) The search for synergy: a critical review from a response surface perspective. Pharmacol Rev 47: 331–385. PMID: 7568331
76. Lee JH, Landrum PF (2006) Development of a multi-component Damage Assessment Model (MDAM) for time-dependent mixture toxicity with toxicokinetic interactions. Environ Sci Technol 40: 1341–1349. PMID: 16572795
77. Schoen ED (1996) Statistical designs in combination toxicology: a matter of choice. Food Chem Toxicol 34: 1059–1065. PMID: 9119316
78. Faust M, Altenburger R, Backhaus T, Blanck H, Boedeker W, et al. (2001) Predicting the joint algal toxicity of multi-component s-triazine mixtures at low-effect concentrations of individual toxicants. Aquat Toxicol 56: 13–32. PMID: 11690628
79. Scheiner SM (1993) Genetics and Evolution of Phenotypic Plasticity. Annual Review of Ecology and Systematics 24: 35–68.
80. Nagano AJ, Sato Y, Mihara M, Antonio BA, Motoyama R, et al. (2012) Deciphering and prediction of transcriptome dynamics under fluctuating field conditions. Cell 151: 1358–1369. doi: 10.1016/j.cell.2012.10.048 PMID: 23217716
81. Via S, Lande R (1985) Genotype-Environment Interaction and the Evolution of Phenotypic Plasticity. Evolution 39: 505–522.
82. Gerard JF, Vancassel M, Laffort B (1993) Spread of phenotypic plasticity or genetic assimilation: the possible role of genetic constraints. J Theor Biol 164: 341–349. PMID: 8246523
83. Schlichting CD, Pigliucci M (1993) Control of phenotypic plasticity via regulatory genes. Am Nat 142: 366–370. doi: 10.1086/285543 PMID: 19425982
84. Wilson M, Lindow SE (1993) Effect of phenotypic plasticity on epiphytic survival and colonization by Pseudomonas syringae. Appl Environ Microbiol 59: 410–416. PMID: 8434910
85. Tonsor SJ, Elnaccash TW, Scheiner SM (2013) Developmental instability is genetically correlated with phenotypic plasticity, constraining heritability, and fitness. Evolution 67: 2923–2935. doi: 10.1111/evo.12175 PMID: 24094343

# Appendix W – Manuscript – 2: $\ell_2$ multiple kernel fuzzy SVM-based data fusion for improving peptide identification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2480084, IEEE/ACM Transactions on Computational Biology and Bioinformatics



# $\ell_2$ multiple kernel fuzzy SVM-based data fusion for improving peptide identification

Ling Jian, Zhonghang Xia, Xinnan Niu, Xijun Liang, Parimal Samir, and Andrew J. Link

**Abstract**—SEQUEST is a database-searching engine, which calculates correlation score between observed spectrum and theoretical spectrum deduced from protein sequences stored in a flat text file, despite it is not a relational and object-oriental repository. Nevertheless the SEQUEST score functions fail to discriminate between true and false PSMs accurately. Some approaches, such as PeptideProphet and Percolator have been proposed to address the task of distinguishing true and false PSMs. However, most of these methods employ time-consuming learning algorithms to validate peptide assignments [1]. In this paper, we propose a fast algorithm for validating peptide identification by incorporating heterogeneous information from SEQUEST scores and peptide digested knowledge. To automate the peptide identification process and incorporate additional information, we employ $\ell_2$ multiple kernel learning (MKL) to implement the current peptide identification task. Results on experimental datasets indicate that compared with state-of-the-art methods, i.e., PeptideProphet and Percolator, our data fusing strategy has comparable performance but reduces the running time significantly.

**Index Terms**—fuzzy SVM, mass spectrometry, multiple kernel learning, peptide-spectrum matches, peptide identification

## 1 Introduction

Efficient mass spectrometry-based (MS) strategies for peptide identification and quantification are prerequisites for performing advanced proteomics studies [2]. Database search engines, such as SEQUEST and MASCOT, have been widely used to automatically match peptide spectra generated from LC/MS/MS experiments to theoretical fragmentation spectra derived from target databases. However, a large number of these peptide scoring matches are false and need to be distinguished from true hits [3].

A number of machine learning approaches have been developed to validate the target PSMs using different scoring functions. PeptideProphet employs the expectation maximization (EM) method to compute conditional probabilities of true PSMs for observed peptide sequences based on the assumption that the true and false PSM data are drawn from a mixture of Gaussian and Gamma distributions [4]. The discrimination scores between true and false PSMs are derived from those conditional probabilities, and all PSMs with scores above a threshold are reported as true PSMs. Choi and Nesvizhskii improve the performance of PeptideProphet using semi-supervised learning [5]. In an alternative approach, Percolator searches against target and decoy databases separately and uses q-values to evaluate the quality of PSMs [6], [7]. Progresses on different post-database searching algorithms have been extensively discussed in [8].

Although a variety of peptide identification algorithms have been proposed, the validated PSMs reported by these algorithms show significant differences [8]. For instance, SVM-based algorithms show that some true PSMs are very close to the decision hyperplane and hinge with decoy PSMs [9], and thus they are difficult to be distinguished from false PSMs. It has been shown that merging different data sources can improve the accuracy of peptide identification [10], [11], [12], [13]. For example, the number of tryptic termini (NTT) of peptides assigned to spectra is valuable information and employed in PeptideProphet [4]. However, using protein information, such as "sibling peptides", has been recoginized improper by researchers [14], [15]. Because that feature may exclude the decoy proteins and cause the target-decoy approach to give biased results.

Although additional scoring and database attributes improved validation of PSMs, data representation for integrating the additional information remains challenges. MKL is an efficient way to combine multiple data sources and has been shown its effectiveness in genomic data fusion [16], visual object detection [17] and face recognition [18].

In this work, we propose a data fusion method based on $\ell_2$ MKL fuzzy SVM (MFS) to integrate multiple data sources for accurate PSM validation. During the course of kernel design, we used SEQUEST searching scores and NTT as attributes. $\ell_2$ MKL model was used to learn the optimal kernel coefficients. In addition, as decoy PSMs are artificially generated, the corresponding labels are known certainly. On the other hand, a large number of target PSMs are false, and thus the labels are not trustworthy either. Hence, these two types of samples should be treated in different confidence. During training process, the fuzzy SVM model assigns a weight to each PSM as its fuzzy membership to reflect its corresponding confident level. Experimental studies show that under FDR equals


- *L. Jian and X. Liang are with College of Science, China University of Petroleum, Qingdao, 266580, China. E-mail: jianlingupc@126.com, liangxijunsd@163.com.*
- *Z. Xia is with the Department of Computer Science, Western Kentucky University, Bowling Green, KY 42101. E-mail: zhonghang.xia@wku.edu.*
- *X. Niu and A.J. Link are with Department of Pathology, Microbiology and Immunology, Vanderbilt University, Nashville, TN 37240. E-mail: {xinnan.niu, andrew.link}@vanderbilt.edu.*
- *P. Samir is with Department of Biochemistry, Vanderbilt University, Nashville, TN 37240. E-mail: parimal.samir@gmail.com.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2480084, IEEE/ACM Transactions on Computational Biology and Bioinformatics



0.05, the MFS method can identify at least 18% more target PSMs than single kernel fuzzy SVM using four different training datasets.

## 2 METHODS

### 2.1 $\ell_2$ MKL of Fuzzy SVM

Different with traditional SVM approaches, which rely on trustworthy labels of training data [19], the MFS model needs to cope with a large number of incorrect training labels. To deal with those untrustworthy labels, Lin and Wang proposed to introduce fuzzy membership during SVM learning [20]. The effectiveness of fuzzy SVM on peptide identification has been shown by Liang *et al.* [9].

Given $l$ PSMs $\mathbf{x}_i \in R^n$, $i=1,\cdots,l$ with class label $y_i \in \{-1,1\}$ and the corresponding fuzzy degree $s_i \in [0,1]$. The kernel-based fuzzy SVM learn model [20] can be written as

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{l} s_i\xi_i \tag{1}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i)+b) \geq 1-\xi_i, i=1,\cdots,l, \tag{2}$$

$$\xi_i \geq 0, i=1,\cdots,l, \tag{3}$$

where $\Phi(\mathbf{x})$ is the mapping from input space to the feature space and $C$ is a regularization parameter, balancing between margin and error.

By solving the dual problem of Eqs.(1-3), one has fuzzy kernel-based classifier

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i k(\mathbf{x}_i,\mathbf{x}) + b, \tag{4}$$

where $\alpha_i, i=1,\cdots,l$ are non-negative Lagrange multipliers and $k(\mathbf{x}_i,\mathbf{x}_j)$ is the kernel function. The performance of a kernel-based classifier closely depends on the design of the kernel matrix which defines pairwise similarity of data points. A good design can capture the most important information hidden in the dataset. MKL provides a general solution to combine multiple single kernels [21], [22], [23], [24]. However, the conventional solution for a MKL model usually degenerates and a single kernel dominates all others. Solutions for non-sparse kernel coefficients were described in [23] and [24] by using $\ell_2$-norm MKL algorithms and showed good performance in the area of bioinformatics [25]. The problem of $\ell_2$-norm MKL for fuzzy SVM can be formulated as follows

$$\min_{\mu} \ \omega\Big(\sum_{i=1}^{p} \mu_i k^i\Big) \tag{5}$$

$$\text{s.t.} \quad \|\mu\|_2 = 1, \tag{6}$$

$$\mu_i \geq 0, i=1,\cdots,p, \tag{7}$$

where $\omega(k)$ denotes the optimal value of the dual problem Eqs.(1-3). To tackle the computational complexity, Sonnenburg *et al.* reformulated the problem as a semi-infinite programming (SIP) [24]. The SIP formulation of MFS can be written as

$$\max \quad \theta \tag{8}$$

$$\text{s.t.} \quad \|\mu\|_2 \leq 1, \tag{9}$$

$$\mu_i \geq 0, i=1,\cdots,p, \tag{10}$$

$$\frac{1}{2}\sum_{i=1}^{p} \mu_i f_i(\alpha) - \sum_{i=1}^{l} \alpha_i \geq \theta, \tag{11}$$

$$0 \leq \alpha_i \leq s_i C, i=1,\cdots,l, \tag{12}$$

$$\sum_{i=1}^{l} y_i\alpha_i = 0, \tag{13}$$

where $f_i(\alpha) = \sum_{j,k=1}^{l} \alpha_j\alpha_k y_j y_k k^i(\mathbf{x}_j,\mathbf{x}_k), i=1,\cdots,p.$

### 2.2 Kernel Design

Several types of data sources are contained in SEQUEST searching results: SEQUEST searching scores, tryptic termini data, etc. Table 1 summarizes the features used by MFS.

TABLE 1
SUMMARY OF FEATURES USED BY MFS FOR SEQUEST SEARCH RESULTS

| Feature | Description |
|---|---|
| Xcorr | the first coss-correlation value from the SEQUEST search |
| ΔCn | the difference correlation value between the first hit and the second hit |
| Sprank | the preliminary score preformed by SEQUSET |
| Ions | the fraction of matched ions |
| Mass | the observed monoisotopic mass of the identified peptide |
| enzN/C | a Boolean value indicating if the peptide has a tryptic N/C-terminus |

We defined two individual kernels and combined them into a comprehensive kernel with the MKL technique.

First, we employed Gaussian kernel to exploit the pairwise similarity between two PSMs based on the SEQUEST search scores. Here, we used five attributes, i.e., Xcorr, ΔCn, Sprank, Ions, and Mass. To avoid attributes with larger values dominating ones with smaller values, we normalized each of the original SEQUEST scores by using the equation *x_nor* = *x_raw* − (mean of *x_raw*)/(*std of x_raw*), where *x_nor* is the normalized SEQUEST score, *x_raw* is the original SEQUEST score, and *std of x_raw* is the standard deviation of the original SEQUEST score. Let $\sigma$ be the number of attributes characterizing the PSM. Then, the first individual kernel is defined by

$$k_{ij}^1 = \exp\Big(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\sigma}\Big). \tag{14}$$

The second kernel was defined to utilize proteolytic information for validating the PSMs. Usually only a very small number of fully-canonical PSMs are assigned to decoy peptides in the database search process [26] and the NTT distributions among true and false peptide assignments are sufficiently distinct. Hence, the NTT of peptide is valuable information in peptide identification [4]. Taking into account the NTT information, the second kernel can be defined as

$$k_{ij}^2 = \exp(-\|n_i-n_j\|^2). \tag{15}$$

Here, $n_i$ stands for the NTT of peptide associated with the $i$th PSM.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2480084, IEEE/ACM Transactions on Computational Biology and Bioinformatics



### 2.3 Fuzzy Membership Design

As true PSMs are usually close to each other, we employed density measurement of target PSMs to construct the fuzzy membership for alleviating the impact of noise. For a given training dataset $\mathbb{D}_{train}=\{(\mathbf{x}_i,y_i)\}_{i=1}^{l}$, denote $l^+$ target PSMs by $\mathbb{D}^{+}_{train}=\{(\mathbf{x}_i^+,1)\}_{i=1}^{l^+}$ and $l^-$ decoy PSMs by $\mathbb{D}^{-}_{train}=\{(\mathbf{x}_i^-,-1)\}_{i=1}^{l^-}$, respectively. The fuzzy membership of decoy PSMs are set as 1 as they are trustworthy. Afterwards the fuzzy membership of target PSMs are normalized to [0 1] by

$$s_i^+=\frac{d(\mathbf{x}_i^+)-\min\limits_{1\le j\le l^+} d(\mathbf{x}_j^+)}{\max\limits_{1\le j\le l^+} d(\mathbf{x}_j^+)-\min\limits_{1\le j\le l^+} d(\mathbf{x}_j^+)}w(\mathbf{x}_i^+) \tag{16}$$

where $w(\mathbf{x}_i^+)$ pre-determined parameters by users, and the density support function $d(\mathbf{x})$ was obtained by solving one-class SVM [27]

$$\min_{\mathbf{w},\rho,\xi}\ \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}+\frac{1}{\nu l}\sum_{i=1}^{l}\xi_i-\rho \tag{17}$$

$$\text{s.t.}\quad \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i)\ge\rho-\xi_i, i=1,\cdots,l, \tag{18}$$

$$\xi_i\ge 0, i=1,\cdots,l. \tag{19}$$

The pseudocode for the MFS method is summarized as follows.

**Algorithm:** $\ell_2$-MFS algorithm

Input: $\mathbb{D}^+,\mathbb{D}^-,\mathbb{D}^{+}_{train},\mathbb{D}^{-}_{train},w,\gamma$

Output: $\mathbb{D}^+$

1: Train one-class SVM decision function $d(\mathbf{x})$ on $\mathbb{D}^{-}_{train}$;
Compute density support $d(\mathbf{x}_i^+)$ of positive samples $\mathbf{x}_i^+\in\mathbb{D}^{+}_{train}$.

2: Compute the fuzzy membership of positive samples $\mathbf{x}_i^+$ via Eq.(16).

3: Calculate the initial kernel matrix $k^1,k^2$ via Eqs.(14-15).

4: Learn kernel coefficients $\mu_1,\mu_2$ over $\mathbb{D}_{train}$ via Eqs.(8-13);
Solve Eqs.(1-3) with comprehensive kernel $k=\mu_1k^1+\mu_2k^2$;
Compute the final decision function Eq.(4)

5: Calculate the decision value $f(\mathbf{x}_i)$ of samples $\mathbf{x}_i\in\mathbb{D}$;
Sort $f(\mathbf{x}_i^-)$ in descending order;
Let $f(\mathbf{x}_\gamma^-)$ be the $\gamma\left|\mathbb{D}^-\right|$th largest decision value in negative samples;
Remove $\mathbf{x}_i^+$ from $\mathbb{D}^+$ if $f(\mathbf{x}_i^+)<f(\mathbf{x}_\gamma^-)$;
Update $\mathbb{D}^+$.

## 3 Results and Discussions

### 3.1 Datasets and Parameter Setting

The LC-MS/MS datasets used in this study were described in a previous publication [26]. In fuzzy SVM, the regularization parameter $C$ is set as 1. In the one-class SVM, we use the default setting for the kernel width $\sigma$ and $\nu$, i.e., set $\sigma$ as the dimension of input variable and $\nu$ as 0.5 [28].

### 3.2 Results on Training Set

The RAW files generated from the different LC/MS/MS experiments were converted to mzXML format using the program ReadW. The MS/MS spectra were extracted from the mzXML file using the program MzXML2Search [29]. Using SEQUEST, the datasets were searched against either S. cerevisiae (SGD-2010) or human Uniprot (uni280910) databases containing target and decoy protein sequences. All decoy protein sequences were created by reversing the target protein sequences. In this study, according to the distribution of Target/Decoy, we random selected the training set of 2500 PSMs from each dataset. Details about the training sets are shown in Table 2 and Table 3.

TABLE 2
SUMMARY OF LC/MS/MS DATASETS AND SEQUEST SEARCH RESULTS

| Sample | Mass Spectrometer | MiPS | Decoy/Total(%) |
|---|---|---|---|
| PBMC$^{\theta}$ | Orbitrap XL | OFF | 35.12 |
| PBMC$^{\theta}$-train | Orbitrap XL | OFF | 35.60 |
| PBMC$^{\theta}$-test | Orbitrap XL | OFF | 35.12 |
| PBMC | Orbitrap Velos | ON | 30.89 |
| PBMC-train | Orbitrap Velos | ON | 30.72 |
| PBMC-test | Orbitrap Velos | ON | 30.88 |
| Tal08 | Orbitrap XL | ON | 39.30 |
| Tal08-train | Orbitrap XL | ON | 39.44 |
| Tal08-test | Orbitrap XL | ON | 39.36 |
| Gcn4 | LCQ | N/A | 54.99 |
| Gcn4-train | LCQ | N/A | 55.08 |
| Gcn4-test | LCQ | N/A | 55.08 |

TABLE 3
SUMMARY OF UNFILTERED, CATEGORIZED PSMS

| Sample | Target | | | Decoy | | |
|---|---|---|---|---|---|---|
| | Full | Half | Non | Full | Half | Non |
| PBMC$^{\theta}$ | 28561 | 17490 | 30344 | 948 | 10033 | 30375 |
| PBMC$^{\theta}$-train | 610 | 349 | 651 | 21 | 201 | 668 |
| PBMC$^{\theta}$-test | 599 | 390 | 633 | 16 | 194 | 668 |
| PBMC | 110404 | 35915 | 62446 | 2520 | 24682 | 65912 |
| PBMC-train | 917 | 286 | 529 | 22 | 196 | 550 |
| PBMC-test | 963 | 276 | 489 | 23 | 190 | 559 |
| Tal08 | 14893 | 6809 | 20520 | 419 | 5877 | 21042 |
| Tal08-train | 547 | 235 | 732 | 17 | 226 | 743 |
| Tal08-test | 537 | 239 | 740 | 14 | 202 | 783 |
| Gcn4 | 1453 | 1210 | 4040 | 106 | 1465 | 6618 |
| Gcn4-train | 471 | 232 | 681 | 21 | 244 | 851 |
| Gcn4-test | 537 | 239 | 740 | 21 | 187 | 776 |

Full denotes fully tryptic, Half denotes half tryptic, and Non denotes non tryptic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2480084, IEEE/ACM Transactions on Computational Biology and Bioinformatics



*3.2.1 Fuzzy SVM vs. SVM* We tested the performance of fuzzy SVM by comparing it with standard SVM over prepared data. As $\mathbb{D}_{train}^{-}$ is trusty data, we used it to determine the density estimator with one-class SVM Eqs.(17-19), and computed the fuzzy membership Eq.(16) of positive samples $\mathbb{D}_{train}^{+}$. State-of-the-art software libsvm is selected to solve the one-class SVM model [28]. The Gaussian kernel defined in Eq.(14) is selected to express the samples similarity in feature space, i.e., $k(\mathbf{x}_i,\mathbf{x}_j)=\Phi(\mathbf{x}_i)^{\mathrm{T}}\Phi(\mathbf{x}_j)$. Consequently, the fuzzy SVM are trained on dataset $\mathbb{D}_{train}$. The classification accuracy of fuzzy SVM on the dataset $\mathbb{D}_{train}^{-}$ are improved significantly from 54.38% to 88.65%, from 55.08% to 85.55%, from 66.53% to 86.11%, and from 86.56% to 95.13% on PBMC$^{\theta}$-train, PBMC-train, Tal08-train and Gcn4-train, respectively.

*3.2.2 Single kernel fuzzy SVM vs. Multiple kernel fuzzy SVM* The first single kernel matrix $k^1$ was defined by the pairwise similarity of PSMs which are represented by SEQUEST search scores. The SVM corresponding to $k^1$ was obtained by solving Eqs.(1-3). In MKL fuzzy SVM, the kernel combination coefficients of $k^1$ and $k^2$ were learned by solving Eqs.(8-13). The optimal kernel coefficients are [0.7505 0.6609], [0.7604 0.6494], [0.7469 0.6649], and [0.7238 0.6900] on four training datasets, i.e., PBMC$^{\theta}$-train, PBMC-train, Tal08-train and Gcn4-train. Comparison experiments on the above mentioned datasets were conducted. The corresponding results are given in Fig. 1 from the first row to the fourth row. The left column of Fig.1 shows the results of single kernel and the right column shows the results of multiple kernels. Under FDR equals 0.05, multiple kernels can identify 29%, 18%, 71%, and 27% more target PSMs than single kernel on PBMC$^{\theta}$-train, PBMC-train, Tal08-train, and Gcn4-train, respectively. This group of experiments demonstrates that multiple kernel can significantly improve the confidence level's distribution of PSMs.

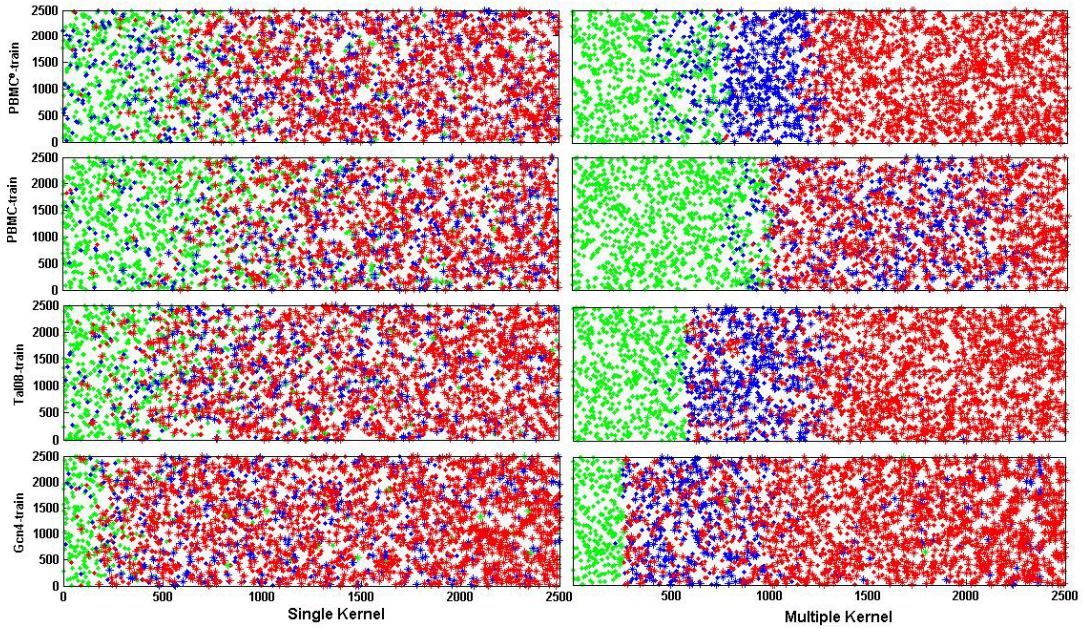



Fig. 1. Ranking results on training datasets: x-coordinate stand for the rank index of samples computed by Eq.(4) and y-coordinate stand for the initial index of samples, circle stand for target points, snow stand for decoy points, green stand for full digested PSMs, blue stand for half digested, red stand for none digested and the vertical line corresponding to FDR equals 0.05.

### 3.3 Performance of $\ell_2$-MFS on Test Datasets

To validate the generalization ability, we also evaluated the performance of the MFS on the test datasets. For each experiment, according to the distribution of Target/Decoy, 2500 PSMs were random drawn from original datasets for testing. The decision function learned in training process is directly used to calculate the decision values of the test data. Details of results are summarized in Fig 2. Note that the distribution of different type points in Fig 2 is closely consistent with the right column of Fig 1. It shows that the performance of MFS is consistent on training and test sets. Moreover, three frequently used criteria in machine learning, i.e., true positive rate (TPR, sensitivity), ture negative rate (TNR), and accuracy (Acc) are listed in Table 4. Highly similar performance in training/test/whole datasets shows that the proposed MFS algorithm can effectively avoid overfitting issue. Hence, the trained model learned in training set can fit the whole dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2480084, IEEE/ACM Transactions on Computational Biology and Bioinformatics



TABLE 4
GENERALIZATION ABILIBYT OF MFS METHOD

| Sample | PBMC[8] | | | PBMC | | | Tal08 | | | Gcn4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TPR(%) | TNR(%) | Acc(%) | TPR(%) | TNR(%) | Acc(%) | TPR(%) | TNR(%) | Acc(%) | TPR(%) | TNR(%) | Acc(%) |
| $\mathbb{D}_{train}$ | 54.60 | 87.19 | 66.20 | 66.80 | 79.56 | 70.72 | 38.77 | 97.57 | 61.96 | 22.62 | 98.98 | 64.68 |
| $\mathbb{D}_{test}$ | 54.25 | 88.04 | 66.12 | 68.98 | 77.59 | 71.64 | 38.06 | 97.56 | 61.48 | 21.46 | 98.84 | 64.08 |
| $\mathbb{D}$ | 53.59 | 86.78 | 66.25 | 66.42 | 77.42 | 69.82 | 37.51 | 97.62 | 61.13 | 22.47 | 98.83 | 64.46 |

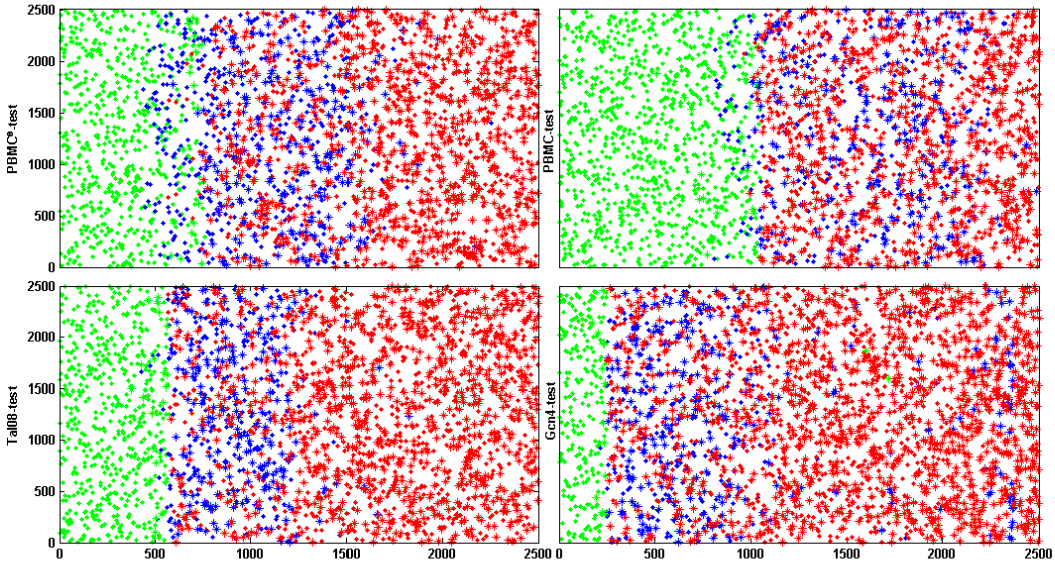



Fig. 2. Ranking results of MFS on test sets.

### 3.4 Evaluating the Performance on the Whole Dataset

Finally, we compared the performance of $\ell_2$-MFS with that of the PeptideProphet and Percolator. For fair comparison our algorithm to PeptideProphet and Percolator, the validation of the PSMs was performed by testing a range of probability filters (p values) provided by PeptidepProphet until the desired FDR (i.e., 0.05) was reached. Likewise, a range of q values were chosen by Percolator until the desired FDR was reached. In the current experiment, the parameters are set as $p \geq 0.25$ for PeptideProphet and $q \leq 0.017$ for Percolator on dataset PBMC[8], $p \geq 0.15$ and $q \leq 0.015$ on PBMC, $p \geq 0.4$ and $q \leq 0.026$ on Tal08, $p \geq 0.25$ and $q \leq 0.024$ on Gcn4. It should be noted that the MKL process is time-consuming, besides storage of the non-sparse kernel matrices is memory cost. So it is difficult for MFS to deal with large scale problem directly. Fortunately, the experiment in Section 3.3 shows that the kernel coefficients and decision function learned in training set can fit the whole dataset. In this section, we use the kernel coefficients and decision function learned in training set to process the whole data set. It takes ~172s to learn the kernel coefficients on training set and ~31s to process the whole PBMC[8] dataset, ~188s to learn the kernel coefficients and ~81s to process the whole PBMC dataset, ~232s to learn the kernel coefficients and ~17s to process the whole Tal08 dataset, and ~141s to learn the kernel coefficients and ~4s to process the whole Gcn4 dataset. Once the decision function is learned in training set, the evaluating process on the whole data set is highly effective. Compared with PeptideProphet and Percolator, the running time can be significantly reduced from hour to minute to process data set with capacity of 100 thousands. The experiments are run on a machine configured with 4 G RAM and core i3-2130 3.4 GHz processor. It should be point out that as the number of half digested PSMs validated by MFS is less than PeptideProphet and Percolator, the total number of validated PSMs under FDR equals 0.05 is less than well-known methods. Table 6 lists detailed validated PSMs according to their digestion pattern: fully-canonical, half-canonical, and non-canonical PSMs. Compared with PeptideProphet, MFS can identify more full digested PSMs on three datasets out of four datasets. Compared with Percolator, MFS can identify more full digested PSMs on two datasets out of four datasets. Hence, the proposed method's performance is comparable with the well-known methods.

Fig. 3 shows the overlap of the identified target PSMs by the three methods on datasets PBMC[8], PBMC, Tal08 and Gcn4. On all the datasets, the target PSMs output by MFS have large overlap with PeptideProphet and Percolator. On PBMC[8], PeptideProphet shared 91.3% target PSMs with MFS; Percolator shared 86.3% target PSMs with MFS. On PBMC, these percentages are 93.3% and 88.0%. On Tal08, these percentages are 94.0% and 91.4%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2480084, IEEE/ACM Transactions on Computational Biology and Bioinformatics



On Gcn4, the percentages are 91.2% and 87.9%, respectively. The results indicate that the majority of target PSMs validated by PeptideProphet and Percolator were also validated by the proposed MFS.

TABLE 5
SUMMARY OF PSMS VALIDATED BY DIFFERENT APPROACHES

| Approach | PBMC* | | PBMC | | Tal08 | | Gcn4 | |
|---|---|---|---|---|---|---|---|---|
| | Target | Decoy | Target | Decoy | Target | Decoy | Target | Decoy |
| PeptideProphet | 35673 | 869 | 120961 | 2947 | 15638 | 387 | 1443 | 38 |
| Percolator | 36096 | 866 | 122568 | 3133 | 14371 | 354 | 1394 | 35 |
| MFS | 33399 | 868 | 115948 | 3165 | 14937 | 382 | 1350 | 36 |

TABLE 6
DISTRIBUTION OF VALIDATED PSMS

| Approach | PBMC* | | | PBMC | | | Tal08 | | | Gcn4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full | Half | Non | Full | Half | Non | Full | Half | Non | Full | Half | Non |
| PeptideProphet | 27622 | 7649 | 402 | 107730 | 13001 | 230 | 14539 | 1088 | 11 | 1375 | 68 | 1 |
| Percolator | 28729 | 7046 | 321 | 111990 | 10453 | 125 | 13855 | 516 | 0 | 1342 | 51 | 1 |
| MFS | 27926 | 5448 | 25 | 109747 | 5728 | 473 | 14721 | 214 | 2 | 1345 | 5 | 0 |

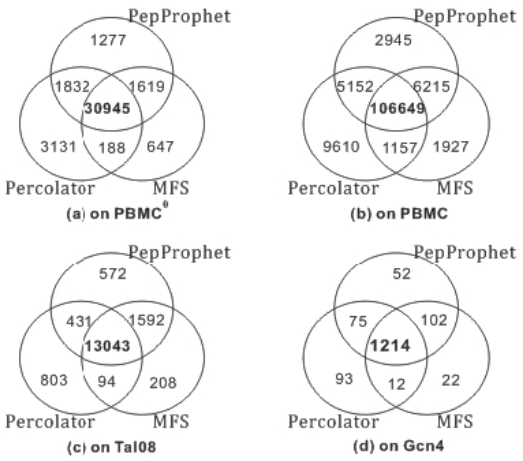



Fig. 3. Overlap of identified target PSMs by PeptideProphet, Percolator and MFS.

## 4 CONCLUSION

In this study, we propose a fast post-database search algorithm $\ell_2$-MFS that can integrate SEQUEST scores and NTT information for validating peptide identification. The main contribution of our approach are: introduce fuzzy membership to effectively suppress the influence of noise; design kernel matrices to represent different kinds of heterogeneous information in unified mathematical format; employ $\ell_2$-norm MKL algorithm to realize the optimal data fusion of the multiple information sources. The MFS algorithm has been assessed by comparing it with the performance of PeptideProphet and Percolator on four data samples, and the validated PSMs demonstrate MFS algorithm provides us with a new way to validate PSMs.

## ACKNOWLEDGMENT

Andrew J. Link and Xinnan Niu were supported by National Institutes of Health under Grant No. GM064779. Ling Jian and Xijun Liang were supported by National Natural Science Foundation of China under Grant No. 11326203 and No. 61403419, Natural Science Foundation of Shandong Province under Grant No. ZR2013FQ034 and ZR2014AP0004, and Fundamental Research Funds for the Central Universities. Data for the project has been funded in part with Federal funds from the National Institutes of Allergy and Infectious Disease, National Institute of Health, Department of Health and Human Service under Contact No. 272200800007C and the Vanderbilt Clinical and Translational Science Award under Grant No. NIH RR024975.

## REFERENCES

[1] F. X. Wu, P. Gagné, A. Droit, G. G. Porier, "RT-PSM, a Real-time Program for Peptide-spectrum Matching with Statistical Significance," *Rapid Communications in Mass Spectrometry*, vol. 20, no.8, pp. 1199-1208, 2006.

[2] F. X. Wu, P. Gagné, A. Droit, G. G. Porier, "Quality Assessment of Peptide Tandem Mass Spectra," *BMC Bioinformatics*, vol. 9, no. 6, pp. 1-10, 2008.

[3] J. Elias and S. Gygi, "Target-decoy Search Strategy for Increased Confidence in Large-scale Protein Identifications by Mass Spectrometry,"*Nature Methods*, vol. 4, no. 3, pp. 207-214, 2007.

[4] A. Keller, A. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search," *Analytical Chemistry*, vol. 74, no. 20, pp. 5383-5392, 2002.

[5] H. Choi and A. I. Nesvizhskii, "Semisupervised Model-based Validation of Peptide Identifications in Mass Spectrometry-based Proteomics," *J. Proteome Res.*, vol. 7 no. 1, pp. 254–265, 2007.

[6] L. Käll, J. Canterbury, J. Weston, W. Noble, and M. MacCoss, "Semi-supervised Learning for Peptide Identification from Shotgun Proteomics Data Sets," *Nature Methods*, vol. 4, no. 11, pp. 923-925, 2007.

[7] P. Yang, J. Ma, P. Wang, *et al.*, "Improving X! Tandem on Peptide Identification from Mass Spectrometry by Self-boosted

Percolator," *IEEE/ACM Trans. Comput. Biology. Bi.*, vol. 9, no. 5, pp. 1273-1280, 2012.

[8] A. I. Nesvizhskii, "A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics," *J. Proteomics*, vol. 73, no. 11, pp. 2092–2123, 2010.

[9] X. Liang, Z. Xia, X. Niu, A. J. Link, L. Pang, F. Wu, and H. Zhang, "Peptide Identification based on Fuzzy Classification and Clustering," *Proteome Sci.*, vol. *11, no.* 1, p. S10, 2013.

[10] K. A. Resing, K. Meyer-Arendt, A. M. Mendoza, L. D. Aveline-Wolf, K. R. Jonscher, K. G. Pierce, W. M. Old, H. T. Cheung, S. Russell, J. L. Wattawa, *et al.*, "Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics," *Anal. Chem.*, vol. 76, no. 13, pp. 3556–3568, 2004.

[11] T. S. Price, M. B. Lucitt, W. Wu, D. J. Austin, A. Pizarro, A. K. Yocum, I. A. Blair, G. A. FitzGerald, and T. Grosser, "Ebp, a Program for Protein Identification Using Multiple Tandem Mass Spectrometry Datasets," *Mol. Cell. Proteomics.*, vol. 6, no. 3, pp. 527–536, 2007.

[12] J. Shi, B. Chen, and F. Wu, "Improve Accuracy of Peptide Identification with Consistency between Peptides," *In Bioinformatics and Biomedicine (BIBM), IEEE International Conference on*, pp. 191–196, 2011.

[13] C. Yang, Z. He, C. Yang, *W. Yu*, "Peptide Reranking with Protein-Peptide Correspondence and Precursor Peak Intensity Information," *IEEE/ACM Trans. Comput. Biology. Bi.*, vol. 9, no. 4, pp. 1212-1219, 2012.

[14] L. J. Everett, C. Bierl, S. R. Master, "Unbiased Statistical Analysis for Multi-stage Proteomic Search Strategies," *Journal of Proteome Research*, vol. 10, no. 4, pp. 2123-2127, 2011.

[15] J. Zhang, L. Xin, B. Shan, et al., "PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification," *Molecular & Cellular Proteomics*, vol. 11, no. 4, doi: 10.1074/mcp.M111.010587, 2012.

[16] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A Statistical Framework for Genomic Data Fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[17] C. Sun and K.-M. Lam, "Multiple-kernel, Multiple-instance Similarity Features for Efficient Visual Object Detection," *IEEE Trans. Image. Process.*, vol. 22, no. 8, pp. 3050–3061, 2013.

[18] Z. Wang and X. Sun, "Multiple Kernel Local Fisher Discriminant Analysis for Face Recognition,"*Signal. Process.*, vol. 93, no. 6, pp. 1496–1509, 2013.

[19] B. Schölkopf, and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press, pp. 227-248, 2002.

[20] C. Lin and S. Wang, "Fuzzy Support Vector Machines", *IEEE Trans. Neural. Network.*, vol. 13, no. 2, pp. 464–471, 2002.

[21] G. R. Lanckriet, Cristianini, P. Bartlett, L. E. Ghaoui, and M. J. Jordan, "Learning the Kernel Matrix with Semidefinite Programming,"*J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.

[22] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple Kernel Learning, ConicDuality, and the SMO Algorithm,"*In Proc. 21th ICLM*, pp. 6., 2004.

[23] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg, "Non-sparse Multiple Kernel Learning,"*In NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008.

[24] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large Scale Multiple Kernel Learning," *J. Mach. Learn. Res.*, vol. **7**, pp. 1531–1565, 2006.

[25] S. Yu, T. Falck, A. Daemen, L. C. Tranchevent, J. A. Suykens, B. De Moor, and Y. Moreau, "L2-norm Multiple Kernel Learning and its Application to Biomedical Data Fusion," *BMC Bioinformatics*, vol. 11, no. 1, pp. 309, 2010.

[26] L. Jian, X. Niu, Z. Xia, P. Samir, C. Sumanasekera, Z. Mu, J. L. Jennings, K. L. Hoek, T. Allos, L. M. Howard, *et al.*, "A Novel Algorithm for Validating Peptide Identification from a Shotgun Proteomics Search Engine," *J. Proteome Res.*, vol. 12, no. 3, pp. 1108–1119, 2013.

[27] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, *et al.*, "Estimating the Support of a High-dimensional Distribution," *Neural. Comput.*, vol. 13, no. 7, pp. 1443-1471, 2001.

[28] C. Chang and C. Lin, "LIBSVM : a Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27, pp. 1-27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[29] E. W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, H. Lam, N. Tasman, Z. Sun, E. Nilsson, B. Pratt, B. Prazen, *et al.*,"A Guided Tour of the Trans-proteomic Pipeline," *Proteomics*, vol. 10, no. 6 ,pp. 1150–1159, 2010.

# Appendix X – Manuscript – 3: A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination

PLOS | ONE

RESEARCH ARTICLE

# A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination

**Kristen L. Hoek[1], Parimal Samir[2], Leigh M. Howard[3], Xinnan Niu[1], Nripesh Prasad[4], Allison Galassie[5], Qi Liu[6], Tara M. Allos[1], Kyle A. Floyd[1], Yan Guo[7], Yu Shyr[8], Shawn E. Levy[4], Sebastian Joyce[1], Kathryn M. Edwards[3]*, Andrew J. Link[1]***

**1** Department of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, TN, 37232, United States of America, **2** Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN, 37232, United States of America, **3** Vanderbilt Vaccine Research Program; Division of Infectious Diseases, Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, TN, 37232, United States of America, **4** HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806, United States of America, **5** Department of Chemistry, Vanderbilt University, Nashville, TN, 27232, United States of America, **6** Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, 37232, United States of America, **7** Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN, 37232, United States of America, **8** Department of Cancer Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, 37232, United States of America

* kathryn.edwards@vanderbilt.edu (KME); andrew.link@vanderbilt.edu (AJL)



OPEN ACCESS

**Citation:** Hoek KL, Samir P, Howard LM, Niu X, Prasad N, Galassie A, et al. (2015) A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination. PLoS ONE 10(2): e0118528. doi:10.1371/journal.pone.0118528

**Academic Editor:** Xu Yu, Massachusetts General Hospital, UNITED STATES

**Received:** August 29, 2014

**Accepted:** December 16, 2014

**Published:** February 23, 2015



**Data Availability Statement:** The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [1] with the dataset identifier PXD001657 and DOI 10.6019/PXD001657. RNA data have been deposited to the GEO database with the dataset identifier GSE64655, (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64655).

**Funding:** This project was funded in part with Federal funds from the National Institutes of Allergy and Infectious Disease, National Institutes of Health,

## Abstract

Systems biology is an approach to comprehensively study complex interactions within a biological system. Most published systems vaccinology studies have utilized whole blood or peripheral blood mononuclear cells (PBMC) to monitor the immune response after vaccination. Because human blood is comprised of multiple hematopoietic cell types, the potential for masking responses of under-represented cell populations is increased when analyzing whole blood or PBMC. To investigate the contribution of individual cell types to the immune response after vaccination, we established a rapid and efficient method to purify human T and B cells, natural killer (NK) cells, myeloid dendritic cells (mDC), monocytes, and neutrophils from fresh venous blood. Purified cells were fractionated and processed in a single day. RNA-Seq and quantitative shotgun proteomics were performed to determine expression profiles for each cell type prior to and after inactivated seasonal influenza vaccination. Our results show that transcriptomic and proteomic profiles generated from purified immune cells differ significantly from PBMC. Differential expression analysis for each immune cell type also shows unique transcriptomic and proteomic expression profiles as well as changing biological networks at early time points after vaccination. This cell type-specific information provides a more comprehensive approach to monitor vaccine responses.

Department of Health and Human Services, under Contract No. 272200800007C, the Vanderbilt Clinical and Translational Science Award grant NIH RR024975, the Childhood Infections Research Program grant T32-AI095202-01, the Immunobiology of Blood and Vascular Systems training grant 5 T32 HL69765-12, and NIH grant GM064779. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Introduction

Systems biology is a comprehensive approach to describe complex interactions between multiple components in a biological system[1]. Using high-dimensional molecular approaches, systems biology identifies changes caused by perturbations such as infection or vaccination, combined with extensive computational analysis to model and predict responses[2,3]. In the context of vaccinology, systems biology offers an approach to dissect the human immune response after immunization by correlating changes in the transcriptome and proteome with antibody or cell-mediated immune responses, in order to make predictions about vaccine efficacy and potentially adverse events[4,5].

The first systems biological studies to dissect human vaccine-induced responses utilized the yellow fever vaccine, YF-17D[6,7]. In these pioneering studies, both CD8+ T cell and B cell signatures identified in microarray profiles were correlated with protective cell-mediated and antibody responses, thus providing predictive signatures. Since these studies, several other vaccines have been studied, including live and inactivated influenza and pneumococcal polysaccharide vaccines [8–10]. Systems biology studies with influenza vaccines identified modules of genes that positively correlated with protective immune responses. For example, interferon-responsive genes that were up-regulated at early time points after TIV vaccination positively correlated with robust hemagglutinin inhibition (HAI) titers[8,10]. Nakaya *et al.* found that an elevated antibody response to trivalent inactivated influenza vaccine (TIV), but not to live attenuated influenza vaccine (LAIV), correlated with upregulation of B cell-specific transcripts, including immunoglobulins (IgA, IgD, IgE, and multiple IgGs) and the TNFRSF17 surface receptor[9]. Using the Nakaya dataset, Tan *et al.* identified immunoglobulin and complement genes as well as proliferation-associated genes to be predictors of protective antibody production in response to TIV vaccination. They concluded that enrichment of these particular gene sets at 7 days post-TIV vaccination was likely due to increased representation of proliferating plasmablasts in subjects with elevated antibody responses[11].

Predictive correlates that can be identified prior to vaccination are emerging in systems vaccinology studies. Tsang *et al.* recently showed that baseline proportions of 126 individual immune cell sub-populations in the blood, identified by comprehensive flow cytometric analysis, could predict influenza vaccine-induced antibody responses[12]. Several studies have found an inverse correlation between baseline influenza-specific microneutralization or HAI titers and the subsequent generation of both plasmablasts and protective antibodies after seasonal influenza vaccination. These studies reported that subjects with lower baseline titers of influenza antibodies generated more robust post-vaccine antibody responses compared to subjects with high baseline titers [12,13]. Furman *et al.* identified several additional baseline predictors of protective immunity, including the frequency of CD8 T cells and NK cells, as well as multiple differentially expressed gene modules. These included genes associated with: 1) apoptotic pathways; 2) cell survival and proliferation (including generation and maintenance of germinal centers); 3) cell-to-cell signaling; 4) RNA post-transcriptional modification; and 5) carbohydrate metabolism [13].

Despite insights into the global human immune responses obtained from these and other studies, the majority of systems biology studies are limited in scope to total RNA from whole blood or peripheral blood mononuclear cells (PBMC)[6–8,11–18]. Since human blood is comprised of a multitude of hematopoietic cell types that are present in varying proportions, responses elicited from under-represented cell types in the blood are likely masked by those of predominant cells[19]. For example, Nakaya *et al.* found upregulation of the transcription factor XBP-1, which is necessary for the terminal differentiation of antibody-forming plasma cells, in RNA from sorted B cells, but not from PBMC, after TIV vaccination[9,20].

Additionally, when utilizing PBMC to monitor the immune response, the contributions of polymorphonuclear (PMN) cells—prime contributors to innate immunity—are overlooked.

Most current vaccines target adaptive immune T and B lymphocytes by conferring lasting, life-long immunity (memory) that can be recalled rapidly upon subsequent encounter with the immunizing antigen[21,22]. The qualitative and quantitative aspects of these adaptive immune responses are slow to develop and are tightly regulated by the rapidly-induced innate immune response. Thus, an immune response represents a highly coordinated effort from multiple hematopoietic cell types—each with their own inherent programming. We therefore believe that it is vitally important to analyze and model individual cell types in response to vaccination.

To develop a comprehensive systems biology model for studying immune responses following vaccination, we developed an efficient protocol to purify from human blood six immune cell types that contribute to both innate and adaptive immune responses: T cells, B cells, natural killer (NK) cells, myeloid dendritic cells (mDC), monocytes, and neutrophils. These cells were isolated and processed immediately for down-stream systems analysis to avoid potential problems associated with the use of frozen cells[23]. Unlike previous systems vaccinology studies, which utilized microarray analysis to map dynamic changes in the transcriptome after vaccination[6–10], this study utilized RNA-Seq data generated prior to and after TIV vaccination together with the human reference genome sequence to identify changes in both protein-coding and non-coding RNA transcripts after vaccination. Additionally, and unique to this study, our protocol included quantitative proteomics to monitor changes in protein expression after vaccination.

Our results reveal that RNA and protein expression profiles from each sorted cell type differ significantly from the profile obtained from PBMC. Comparison of differentially expressed transcripts and proteins after vaccination with 2011–2012 seasonal TIV further shows considerable differences between PBMC and sorted cells. Together, our data suggest that important cell type-specific information is gained when purified cells rather than PBMC or whole blood are utilized in systems studies. The cell type-specific information obtained from unbiased RNA-seq and quantitative proteomics analysis utilizing the complete human reference genome sequence provides a more comprehensive systems biology approach to monitor and eventually to model vaccine responses. This approach is applicable for other systems biology studies involving complex interactions between different cell types following vaccinations, infectious diseases, diseases and pharmacological interventions.

## Materials and Methods

### Seasonal TIV Vaccination of human volunteers and blood collection

Volunteer recruitment and vaccination protocols for this study were approved by the Vanderbilt Institutional Review Board (IRB#111030 "CLR-03 2011-Immune Cells and Soluble Factors from Healthy Donor"). After obtaining written informed consent, thirty one subjects were enrolled in this study. Twenty-nine subjects provided 90mL blood samples to develop our phenotyping and cell sorting protocols and to establish baseline blood profiling information; for these purposes, twenty three subjects provided a single blood sample, and six subjects provided four samples over subsequent days on the same schedule as proposed for vaccinated subjects. Once the cell sorting pipeline was in place, two subjects were vaccinated with a single dose of 2011–2012 seasonal trivalent inactivated influenza vaccine (TIV) (strains included: A/California/7/09 (H1N1,), A/Perth /16/2009 (H3N2), and B/Brisbane/60/2008). Blood samples (90mL) from the two vaccinated subjects were processed prior to vaccination (day 0) and at days 1, 3, and 7 post-vaccination for downstream RNA-seq and quantitative proteomics analysis.

## Immune cell purification and flow cytometric analysis

PBMC and PMN were isolated from anti-coagulated (EDTA) whole blood via Ficoll-paque PLUS (GE Healthcare) separation. Residual RBCs were removed from the PMN fraction by ammonium-chloride-potassium (ACK) lysis (KD Medical). Single cell suspensions of PBMC or PMN were subjected to magnetic bead separation. T cells, monocytes, and neutrophils were enriched by positive selection using directly conjugated anti-CD3, anti-CD14, and anti-CD15 microbeads (Miltenyi Biotec), respectively. B cells were enriched by positive selection using anti-PE beads after staining with anti-CD19-PE antibody (Miltenyi Biotech) since directly conjugated CD19-microbeads interfered with subsequent anti-CD19 phenotypic staining. NK/mDC were enriched by negative selection using Streptavadin microbeads (Miltenyi Biotec) after staining with biotinylated anti-CD19 (clone HIB19), anti-CD15 (clone HI98), anti-CD14 (clone 61D3), and anti-CD3 (clone UCHT1) antibodies (eBioscience). MACS enriched cells were stained with 7-aminoactinomycin D (7-AAD), CD11c-FITC (clone B-ly6) CD15-APC (clone HI98) and CD56-PE-Cy7 (clone B159) (BD Biosciences), as well as CD19-PE (130-091-247), CD3-VioBlue (130-094-363), and CD14-VioGreen (130-096-875) (Miltenyi Biotec), and were subjected to FACS on a BD FACSAriaIII flow cytometer. Cell purity of $\geq$98% was confirmed by re-analysis on the FACSAriaIII after the sort. Whole blood, PBMC, PMN and pooled sorted cells were subjected to 9-color flow cytometric analysis (FCM) to assess phenotype and cellular activation at each time point using the same sorting markers as above, without 7-AAD, and with addition of CD86-PerCP-Cy5.5 (clone FUN-1), CD69-APC-Cy7 (clone FN50), and CD134-PE-Cy5 (clone ACT35) (BD Biosciences). The SPHERO Ultra Rainbow calibration kit (Spherotech; URCP-50-2K) was utilized to control for daily fluctuations in the detectors used for activation marker staining. FCM was performed on a BD LSRFortessa flow cytometer, and data was analyzed using the *FlowJo* software package (Tree Star).

## RNA expression analysis

Total RNA was extracted from PBMC and sorted immune cells ($\leq$0.5x10$^6$ cells) from the two TIV-vaccinated subjects using the automated Maxwell 16 magnetic particle processor and a Maxwell 16 LEV simply RNA kit (Promega Corp.). RNA was quantified by either a Qubit fluorometer (Life Technologies) or the Quant-iT RiboGreen RNA Assay (Life Technologies). To assess RNA integrity, total RNA was evaluated on a Bioanalyzer 2100 (Agilent Technologies). One hundred ng of total RNA with RIN values >7 was required for proceeding to downstream RNA-seq applications. Polyadenylated RNAs were isolated using NEBNext magnetic oligo d $(T)_{25}$ beads. NEBNext mRNA Library Prep Reagent Set for Illumina (New England BioLabs Inc.) was used to prepare individually bar-coded next generation sequencing expression libraries. Library quality was assessed by Qubit 2.0 Fluorometer (Invitrogen), and library concentration was estimated by utilizing a DNA 1000 chip on an Agilent 2100 Bioanalyzer (Applied Biosystems). Accurate quantification of the prepared libraries for sequencing applications was determined using the qPCR-based KAPA Biosystems Library Quantification kit (Kapa Biosystems, Inc.). Each library was diluted to a final concentration of 12.5nM and pooled equimolar prior to clustering. Paired-End (PE) sequencing (25 million, 50-bp, paired-end reads) was performed using a 200 cycle TruSeq SBS HS v3 kit on an Illumina HiSeq2000 sequencer (Illumina, Inc.). Image analysis and base calling was performed using the standard Illumina Pipeline consisting of Real time Analysis (RTA) version v1.13. Raw reads were de-multiplexed using a bcl2fastq conversion software v1.8.3 (Illumina, Inc.) with default settings. Post-processing of the sequencing reads from RNA-seq experiments from each sample was performed as per HudsonAlpha's unique in-house pipeline. Briefly, quality control checks on raw sequence data from each sample were performed using *FastQC* (Babraham Bioinformatics). Raw reads were

mapped to the reference human genome hg19/GRCh37 using TopHat v1.4[24,25]. The alignment metrics of the mapped reads were estimated using *SAMtools* (S1 Dataset. RNA-seq quality control)[26]. Aligned reads were imported onto the commercial data analysis platform *AvadisNGS* v1.5 (Strand Life Sciences). After quality inspection, the aligned reads were filtered on the basis of read quality metrics where reads with a base quality score less than 30, alignment score less than 95, and mapping quality less than 40 were removed. Remaining reads were then filtered on the basis of their read statistics, where missing mates, translocated, unaligned and flipped reads were removed. The reads list was then filtered to remove duplicates. Samples were grouped and quantification of transcript abundance was performed on this final read list using Trimmed Means of M-values (TMM) as the normalization method [27]. Output data utilized for all subsequent comparisons was a normalized signal value generated by *AvadisNGS*. Resulting transcript lists were quality checked using *AvadisNGS* on a cell-type and donor basis across time points using comparative analysis; transcripts from the same cell type and donor required a correlation coefficient >0.9 to be accepted for further analysis (S1 Fig.. RNA quality control).

## Quantitative proteomic analysis

Protein extracts from PBMC and sorted immune cells ($1x10^6$ cells) from the two vaccinated subjects were prepared as previously described[28] using a modified lysis buffer (50% Trifluoroethanol 50 mM HEPES) and quantified by BCA assay[29]. An immune cell common standard (ICCS) control sample composed of protein extracts from PBMC and $CD15^+$ cells (80% and 20%, respectively, by protein weight) was included in all 8plex iTRAQ experiments. Ten ug of reduced, alkylated, and trypsinized protein extracts were labeled with iTRAQ tags (AB Sciex), pooled, and analyzed by MudPIT using an Eksigent 2-D nanoLC pump coupled to a nanoESI-LTQ-OrbitrapXL mass spectrometer (Thermo Scientific)[30,31]. The precursor ions were analyzed in the Orbitrap followed by 4 collision induced dissociation (CID) fragment ion scans in the ion trap to identify peptides. The precursor ions were then fragmented by higher-energy collisional dissociation (HCD) to measure reporter ion intensities in the Orbitrap. For each precursor ion, the CID and HCD spectra were merged using *Proteome Discoverer v1.3* (Thermo Scientific). The merged fragmentation spectra were searched against a forward and reverse concatenated human Ensembl protein and common contaminants database (gene model 74) using the *Sequest* database search engine running under *Proteome Discoverer* [32,33]. Precursor mass tolerance was set to 20 ppm and fragment mass tolerance was set to 0.8 Da. iTRAQ modification of N-terminus and ε-amine of lysines and β-methylthiolation of cysteines were used as static/constant modifications of the peptides. Oxidation of methionine and tryptophan and deamidation of asparagine and glutamine were used as dynamic/variable modifications of the peptides. Protein assembly, reporter ion quantitation and statistical analysis were performed with a 5% peptide and protein FDR using *ProteoIQ* v2.61 (Premier Biosoft). A slope of the regression line >0.8 between the technical replicates of the common control (ICCS) based upon pseudospectral counts was required as a quality control threshold (S2 Fig.. Proteomics quality control).

## Comparative and differential analysis

Comparative analysis of RNA transcripts and proteomics profiles between cell types was performed using Spearman correlation coefficients. Principal component analysis (PCA) was performed in R and plotted using the *rgl* package[34]. Heirarchical clustering analysis and dendograms were generated using *Cluster3.0* and *Java Treeview*, respectively [35,36]. Differential RNA transcript expression analysis was performed in *AvadisNGS v1.5*. RNA transcripts

were first filtered to include only reads that met a threshold of 0.5 RPKM in at least one time point on a per-cell type and per-subject basis. Next, a Z-test (theoretical estimate of variance), in which the Benjamini-Hochberg procedure was used to fix the FDR at 0.05, was applied to pair-wise comparisons (days 0–1, 0–3, and 0–7) on a per cell-type and per-subject basis (*AvadisNSG v1.5*, Strand Life Sciences) [37]. Differential expression of transcripts was then calculated on the basis of fold change[38]. A $\geq$1.5 fold change in expression between time points was considered significant. Venn diagrams were used to identify differentially expressed transcripts between individuals and cell types. To identify potential differential splicing events in the RNA-Seq data, the publically accessible data analysis package *Multivariate Analysis of Transcript Splicing (MATS)* was used[39]. *MATS* uses a multivariate uniform distribution to model the between-sample correlation in exon splicing patterns, and a Markov chain Monte Carlo (MCMC) method coupled with a simulation-based adaptive sampling procedure to calculate the P value and false discovery rate (FDR) of differential alternative splicing. Transcripts expressing the same differential splice event with both a $p \leq 0.05$ and FDR$\leq$0.05 from both subjects were identifeid as significant. For differential protein expression analysis following vaccination, fold changes were calculated in *ProteoIQ*. A plot of log2 fold changes against pseudospectral counts was used to assess the effect of sampling over the observed fold changes. The symmetric distribution of log2 fold changes versus pseudospectral counts suggests the differential expression analysis was unbiased by protein abundances (S2 Fig.). Distribution of fold changes across different samples was visualized using cluster dot plots (S2 Fig.). Missing values and contaminating keratin proteins were removed prior to differential analysis. A $\geq$1.25 fold change in expression between pair-wise comparisons (days 0–1, 0–3, and 0–7) was considered significant. A Unix bash shell command was used to identify differentially expressed proteins shared between individuals and cell types, as well as to create lists of DE genes and proteins for heat maps. Heat maps of RNA and protein fold changes following vaccination were generated using *Cluster3.0* and *Java Treeview*.

### Visualization of RNA and proteins across the human genome

Genome-wide visualization of relative RNA or protein expression from PBMC and each purified immune cell type was generated using the open-source Circos software package[40]. The genome location for individual transcript and protein data points was mapped using *BioMart* [41].

### Network analysis

Differentially expressed protein-coding RNA transcripts and proteins identified in both subjects after vaccination were imported into *Ingenuity Pathway Analysis* (Qiagen) to identify the most significantly affected unique canonical pathways, biological functions and networks between time points.

## Results

### Immune cell isolation

To obtain purified human immune cells, PBMC and PMN were immediately fractionated from freshly collected venous blood over a Ficoll density gradient. Average numbers of PBMC and PMN obtained from 90mL of fresh blood were 232.9 ± 96.6x$10^6$ and 113.1 ± 70.0x$10^6$ (average ± SD), respectively. These cells were stained with a cocktail of antibodies to identify and quantify six targeted immune cell types: CD3+ T cells, CD14+ monocytes, CD15+ neutrophils, CD19+ B cells, CD11c+ mDC, and CD56+ NK cells (Fig. 1A). Distribution of leukocyte cell

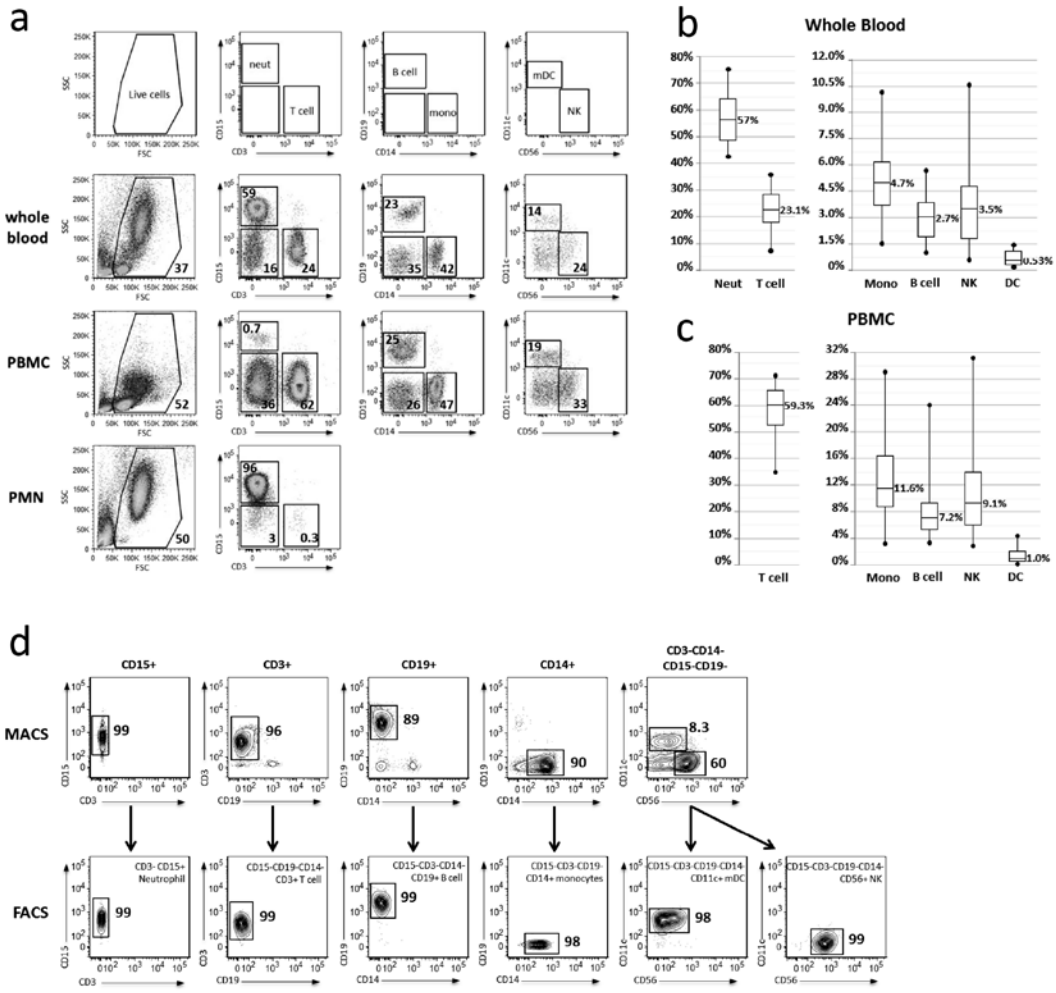



**Fig 1. Flow cytometric analysis of immune cell types purified from human blood. (a)** Whole blood (top panel), PBMC (middle panel) and PMN (bottom panel) cell samples from a single representative subject were stained with a cocktail of antibodies directed against CD3, CD11c, CD14, CD15, CD19, and CD56 cell surface markers for phenotypic analysis by flow cytometry. Moving left-to-right, live cells were first gated for CD3 and CD15 expression. Subsequent gates were drawn from the negative population in the previous panel. **(b,c)** Graphical representation of flow cytometric analysis from whole blood and the PBMC fraction reveals the variability among subjects (n = 31). **(d)** PMN and PBMC cell fractions from a single representative subject were subjected to $CD15^+$, $CD3^+$, $CD19^+$, and $CD14^+$ positive selection or $CD19^-CD15^-CD14^-CD3^-$ enrichment (top panels) via magnetic sorting (MACS). MACS-enriched cells were stained with the same cocktail of antibodies as in (a), with addition of 7AAD to exclude non-viable cells, and subjected to FACS (bottom panels) following the same gating scheme as in (a) to obtain highly purified neutrophil, T cell, B cell, monocyte, m DC, and NK populations for systems analysis.

doi:10.1371/journal.pone.0118528.g001

types in the whole blood (Fig. 1B) and PBMC fraction (Fig. 1C) fell within the expected, physiologically accepted range (whole blood: neutrophils, 25–80%; T cells, 10–30%; B cells, 1–9%; monocytes, 5–11%; NK, 1–8%; mDC 0–1%); however, variability was observed between subjects.

The number of cells needed for enrichment of each cell type, as well as the order of enrichment, depended upon both the total number of PBMC obtained and the individual's phenotypic blood profile. The standard sorting protocol was designed for use when 150–300x$10^6$ PBMC were obtained from 90mL fresh blood, which occurred in 24/39 (62%) samples (S3 Fig. Flow chart for immune cell purification). To account for variability in the abundance and composition of each donor's cells, alternative sorting schemes were developed to maximize recovery of all cell types if larger or smaller numbers of PBMC were obtained from 90mL fresh blood, which occurred in 9/39 (23%) and 6/39 (15%) samples, respectively (S3 Fig.). Additionally, if a phenotypic blood profile varied widely from the average, or if recovery of a particular cell type was sub-optimal on the first visit, the proportion of PBMC or PMN fraction dedicated to enrichment of the affected cell-type(s) was altered accordingly in subsequent visits.

PBMC and PMN fractions were first subjected to magnetic-activated cell sorting (MACS) to positively select CD3+ T cells, CD14+ monocytes, CD15+ neutrophils, CD19+ B cells or negatively enrich for CD3-CD14-CD15-CD19- NK and mDC (Fig. 1D, top panels). However, cell yields and purity were inconsistent, rarely resulting in greater than 90% purity from any sample. Therefore, MACS-enriched cells were further subjected to fluorescence-activated cell sorting (FACS). Using the same antibody cocktail employed for phenotyping, with addition of (7-AAD) to exclude non-viable cells, neutrophils ($CD3^-CD15^+$), T cells ($CD15^-CD19^-CD14^-CD3^+$), B cells ($CD15^-CD3^-CD14^-CD19^+$), monocytes ($CD15^{-/lo}CD3^-CD19^-CD14^+$), mDC ($CD15^-CD3^-CD19^-CD14^-CD56^-CD11c^+$), and NK cells ($CD15^-CD3^-CD19^-CD14^-CD11c^-CD56^+$) were sorted with greater than 98% purity (Fig. 1D, bottom panels) in a short period of time; each sort generally took 30 min or less. Purified cells were not significantly activated by the sorting process, as assessed by flow cytometric analysis of size and scatter as well as surface staining for activation markers (S4 Fig. Individual cell types are not activated by the sorting process).

By employing this approach for sorting 6 immune cells types from fresh whole blood, we consistently obtained sufficient cells for both transcriptomic and proteomics analysis. After FACS purification, cells were immediately processed and frozen for downstream RNA ($\leq$0.5x$10^6$ cells) and protein (1x$10^6$ cells) analyses. Greater than 1.5x$10^6$ of each cell type was

**Table 1. Recovery of purified immune cells.**

| | Starting quantity of PBMC* or PMN# (mean ± SD x $10^6$) | Cell recovery after MACS + FACS (mean ± SD x $10^6$) | N |
|---|---|---|---|
| **T cell** | 21.7 ± 3.3 | 2.0 ± 0.70 | 36 |
| **B cell** | 83.1 ± 20.7 | 1.7 ± 0.69 | 38 |
| **Monocyte** | 75.0 ± 20.7 | 2.9 ± 0.39 | 22 |
| **mDC** | 114.9 ± 38.2 | 0.41 ± 0.31 | 26 |
| **NK** | | 1.9 ± 0.98 | 26 |
| **Neutrophil** | 31.6 ± 8.2 | 2.9 ± 0.35 | 16 |

*T cells, B cells, monocytes, mDC and NK cells were enriched/purified from the PBMC fraction
# Neutrophils were enriched/purified from the PMN fraction.

doi:10.1371/journal.pone.0118528.t001

typically collected, except for mDC (Table 1). Recovery of sorted mDC was sufficient only for RNA analysis; proteomic analysis was not performed on this cell type.

## Transcriptomic and Proteomic analysis in two TIV-vaccinated subjects

Previous systems biology approaches investigating yellow fever and influenza vaccine responses utilized microarray analysis to map the transcriptome after vaccination[6–10]. We used a more comprehensive, sensitive, quantitative and unbiased approach, next-generation RNA sequencing (RNA-Seq), which measures the RNA expression profile of each sample more accurately over a greater dynamic range than microarray-based technologies[42]. In addition to identification of expected coding sequences, RNA-seq allows for identification of non-coding transcripts, splice variants, sequence polymorphisms, and previously unannotated genes [43]. Additionally, the majority of systems vaccinology studies have focused solely on transcriptional analysis to map the immune response, with only selected proteins validated. We also used unbiased quantitative proteomics in addition to transcriptional data to analyze the immune response after vaccination.

A minimum of 100 ng total RNA of high quality (RIN greater than 7) was required for the construction of polyadenylated RNA-seq libraries. Sufficient RNA (250–700 ng total RNA) of good quality was obtained from $0.5x10^6$ PBMC and FACS-sorted T cells, B cells, NK, monocytes and neutrophils, as well as from $0.4–0.5x10^6$ FACS-sorted mDC (S5 Fig. Adequate RNA quality and quantity is obtained from sorted immune cells for RNA-seq applications). While sufficient quantity of RNA was obtained from $0.5x10^6$ neutrophils for our studies, these cells consistently yielded less RNA compared to other cell types, suggesting that additional sorted neutrophils should be collected in the future for downstream RNA applications.

Using 25 million, 50-bp paired-end (PE) RNA-sequencing, the transcriptomes of PBMC as well as the six purified immune cell types from two subjects prior to (day 0) and at days 1, 3, and 7 after TIV vaccination were profiled. After the sequenced reads were aligned to the hg19 human reference genome and filtered to remove transcripts of poor quality, samples were loaded into *AvadisNGS v1.5* for downstream analysis. Approximately 56,000 transcripts were identified in 56 RNA samples (S2 Dataset. Normalized transcript expression in human immune cells prior to and post-TIV vaccination). Of these transcripts, 19,000–27,000 transcripts per cell type contained normalized signal values that were greater than zero. Twenty nine classes of RNA transcripts were identified, including protein coding RNA, pseudogenes, anti-sense RNA, long intervening non-coding RNA (lincRNA), and novel genes (S1 Table. Summary of baseline RNA transcripts identified in each cell type from one subject by RNA-seq analysis). Identification of non-polyA classes of RNA was likely caused by non-specific binding to oligo-dT or other inefficiencies during library construction; however, these classes constituted less than 2% of the total transcripts identified. Using Circos[40], PBMC and purified immune cell baseline (day 0) transcripts from a vaccinated subject plotted over the length of the human genome showed transcription was active across most of the genome, with small regions that appeared transcriptionally silent (S6 Fig. Transcriptional profiling of PBMC and individual immune cell types). Each of the purified immune cell types displayed distinct RNA expression profiles compared to PBMC and the other cell types. Pair-wise comparison of baseline (day 0) transcriptomes from the subject showed weak correlation between PBMC and each sorted cell type (Fig. 2A). Principal component analysis (PCA) of transcriptomes from each time point revealed that all cell types clustered distinctly based on RNA expression profiles (Fig. 2B). Finally, hierarchical clustering analysis of filtered transcripts revealed that each cell type displayed a distinct RNA expression profile that differed from both PBMC and the other cell types in all

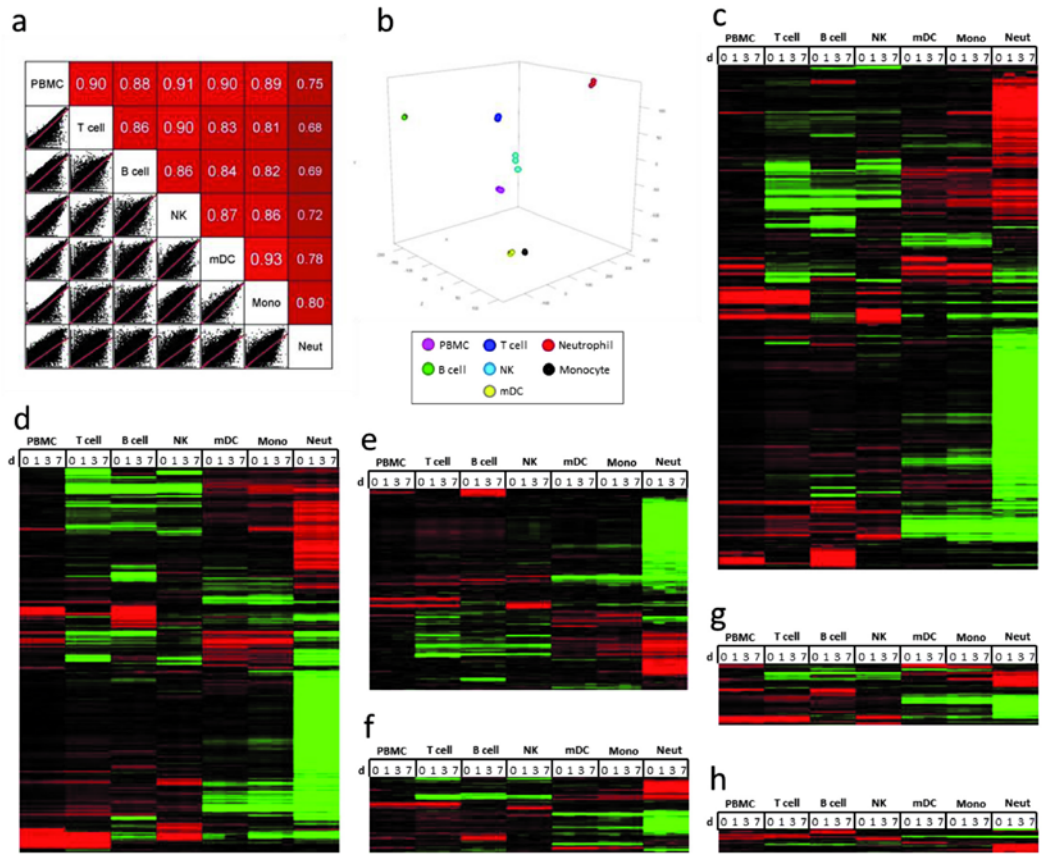



**Fig 2. RNA-Seq analysis of purified immune cells after TIV vaccination. (a)** Pair-wise comparison of day 0 RNA profiles (all transcript classes represented, filtered to remove zero values; 32,505 transcripts) from a vaccinated subject shows that the transcriptome of each sorted cell type correlates weakly with PBMC and other sorted cell types. **(b)** PCA of RNA profiles (all transcript classes represented, filtered to remove zero values; 37,606 transcripts) from a TIV-vaccinated subject at four time points shows that the purified immune cell types cluster into distinct groups, although monocytes and mDC cluster closely. **(c-h)** Semi-supervised hierarchical clustering analysis of RNA expression from a vaccinated individual reveals that purified immune cells have distinct RNA expression profiles compared to PBMC at all time-points. Data (non-zero transcripts with an RPKM of 1 in at least one sample) was centered for normalized signal value across gene and cell type; red = up, black = no change, green = down. **(c)** All transcript classes (21,438 transcripts). **(d)** Protein coding transcripts (13,243 transcripts, including Ig and TCR transcripts). **(e)** Pseudogenes (3,466 transcripts, 2x scale). **(f)** Anti-sense RNA (1,310 transcripts, 2x scale). **(g)** lincRNA (1,047 transcripts, 2x scale). **(h)** New genes (167 transcripts, 5x scale).

doi:10.1371/journal.pone.0118528.g002

classes of RNA investigated (Fig. 2C-H) (S3 Dataset. Normalized transcript expression in human immune cells filtered for an RPKM of 1.0 in at least one sample from one subject).

Prior to performing quantitative proteomics, protein lysates were quantified. PBMC and sorted immune cells ($1x10^6$) generated between 30–80 μg of protein/sample (S7 Fig. Adequate protein quantity is obtained from sorted immune cells for proteomics applications). In contrast to the RNA levels, neutrophils contained the highest amount of protein, while lymphocytes contained the least. Lysates from each sample were trypsinized, desalted, and labeled with

8plex iTRAQ reagents. A control sample—the Immune Cell Common Standard (ICCS)—was labeled with two iTRAQ channels to assess technical variation and used to normalize data across experiments. Two labeling strategies were tested to determine the optimal pooling strategy for detecting proteomic changes after vaccination (S8 Fig. Two iTRAQ strategies for quantitative proteomic analysis of immune cells after vaccination). In strategy 1, all six cell types at a single time point were multiplexed in one experiment. The advantage of this approach is that technical experimental variation between cell types at each time point would be minimized. However, since liquid chromatography tandem mass spectrometry (LC-MS/MS) selected proteins for identification and quantification based upon their abundance in the sample, proteins present in higher amounts across the samples would be preferentially quantified. Thus, differentially changing proteins with low expression from a single cell-type might not be quantified. Also, by increasing the complexity of the sample pool through multiplexing lysates from six different cell types, co-fragmentation of co-eluting peptides might cause an increase in iTRAQ signal interference. In strategy 2, all four time-points from one cell type were multiplexed in a single experiment. The advantage of this approach is that by pooling similar proteomes, sample complexity is reduced, thus reducing iTRAQ signal interference caused by co-fragmentation of co-eluting peptides. Since LC-MS/MS quantifies only a fraction of the proteome, this strategy would also ensure quantification of a larger fraction of cell type-specific proteins. However, cell type-specific changes that are artifacts might be detected due to technical experimental variation. We tested both strategies and analyzed the results using both unsupervised hierarchical clustering and PCA (S8 Fig.). Strategy 2 produced cell-type specific clustering and protein expression patterns by both hierarchical clustering and PCA, while strategy 1 did not. Since the samples in the iTRAQ experiments using strategy 1 did not cluster together by either hierarchical clustering or PCA, we discounted the possibility of batch effect. Therefore, strategy 2 was considered the optimal approach and employed for proteomic analysis.

Peptide spectra generated by LC-MS/MS were searched against the human Ensembl database of protein sequences using *Sequest* [33], and the resulting peptides were scored and assembled into proteins and quantified based upon the iTRAQ reporter ion intensities in *ProteoIQ*. The proteomes of PBMC and five purified immune cell types from two subjects prior to (day 0) and at days 1, 3, and 7 after TIV vaccination were analyzed. Approximately 7,000 proteins were identified in 44 protein samples (S4 Dataset. Normalized protein expression in human immune cells prior to and post-TIV vaccination). After removing zero values and contaminating keratins, approximately 4,000 proteins from each subject were retained for further analysis (S5 Dataset. Normalized protein expression in human immune cells filtered to remove zero values and contaminating keratins from one subject). Similar to transcriptomic analysis, the PBMC and purified immune cell baseline (day 0) proteomes from a vaccinated subject plotted over the length of the human genome showed activity across the majority of the genome (S9 Fig. Proteomic profiling of PBMC and individual immune cell types). Additionally, each of the purified immune cell types displayed distinct proteomic profiles when compared to PBMC and the other cell types. Pair-wise comparison of baseline (day 0) proteomic data from the subject showed poor correlation between PBMC and sorted cell types (Fig. 3A). PCA of proteomic data from each time point revealed that all cell types clustered distinctly based on proteomic profiles (Fig. 3B). Hierarchical clustering analysis of proteins identified showed that each cell type displayed a distinct protein expression profile that differed from both PBMC and the other cell types (Fig. 3C).

Strikingly, when clustering samples from both subjects in the same experiment by PCA, cell types from both subjects at every time point clustered similarly for RNA expression ( ~ 39,000 transcripts, filtered to remove zero values). However, when analyzing protein data from both subjects ( ~ 5,300 proteins, filtered to remove zero values and contaminating keratins), samples

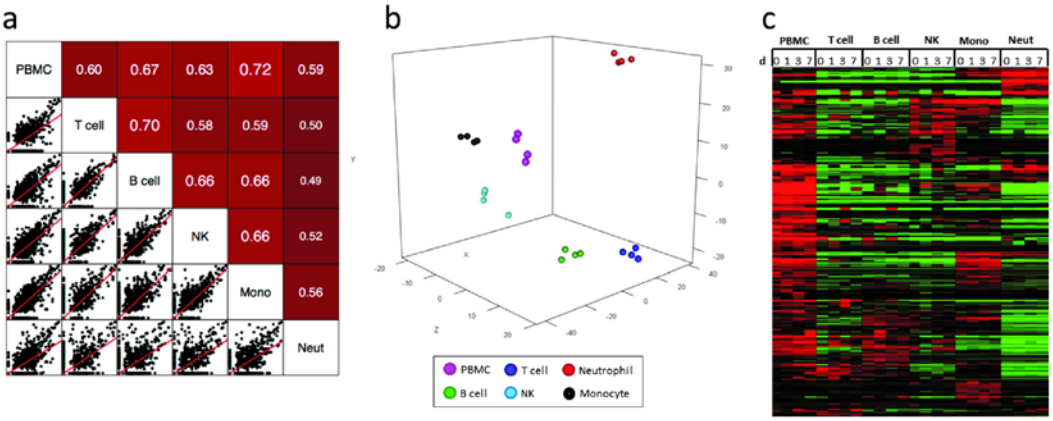



**Fig 3. Proteomic analysis of purified immune cells after TIV vaccination. (a)** Pair-wise comparison of day 0 protein profiles (3,852 proteins, filtered to remove zero values and contaminating keratins) from a vaccinated subject shows that proteomes of sorted cells correlate poorly with PBMC. **(b)** PCA of protein profiles from a TIV-vaccinated subject at four time points shows that purified immune cell types cluster into distinct groups. **(c)** Semi-supervised hierarchical clustering analysis of relative protein expression from a vaccinated individual reveals that purified immune cells have distinct proteomic expression profiles compared to PBMC. Data was centered across protein and cell type; red = up, black = no change, green = down.

doi:10.1371/journal.pone.0118528.g003

of the same cell type at every time point clustered similarly on a per-subject basis, but cells from the two subjects did not cluster together (S10 Fig. Principal component analysis reveals poor correlation of proteomes between subjects).

## Differential analysis of RNA and proteins from two TIV-vaccinated subjects

For comparison of transcriptional changes in PBMC and sorted immune cells, transcripts that were differentially expressed (DE) ≥1.5-fold ($p \leq 0.05$) after vaccination were investigated. While standard methods for determining fold change typically use a 2x fold-change, we found that using this threshold failed to identify significant numbers of shared DE transcripts between both subjects. We therefore tested several different fold-change values, ranging from 1.25x-1.75x. By lowering the threshold to 1.5x, we obtained more comprehensive lists of DE transcripts from each cell type that were shared between both donors at each time point. When DE transcripts from PBMC were compared to DE transcripts from each purified immune cell type, less than 10% similarity was typically observed (S2 Table. Comparison of differentially expressed RNA transcripts in PBMC and individual immune cell types). Circos was used to plot DE transcripts from PBMC and each purified immune cell type from a vaccinated subject over the length of the human genome and to visualize overlap of differentially expressed genes at three time points after TIV vaccination (day 1, day 3, and day 7) (Fig. 4). The plots showed a lack of substantial overlap in differential expression between PBMC and each purified immune cell type. Interestingly, the three time points showed changing patterns of overlapping expression for PBMC and each cell type after TIV vaccination. Substantial variability was also observed in the number of cell type-specific DE transcripts when making subject-to-subject comparisons, with less than 10% similarity between donors for most cell types and time points (S3 Table. Shared DE RNA transcripts; S6 Dataset. Shared up-regulated DE RNA transcripts; and S7 Dataset. Shared down-regulated DE RNA transcripts). To minimize background noise,

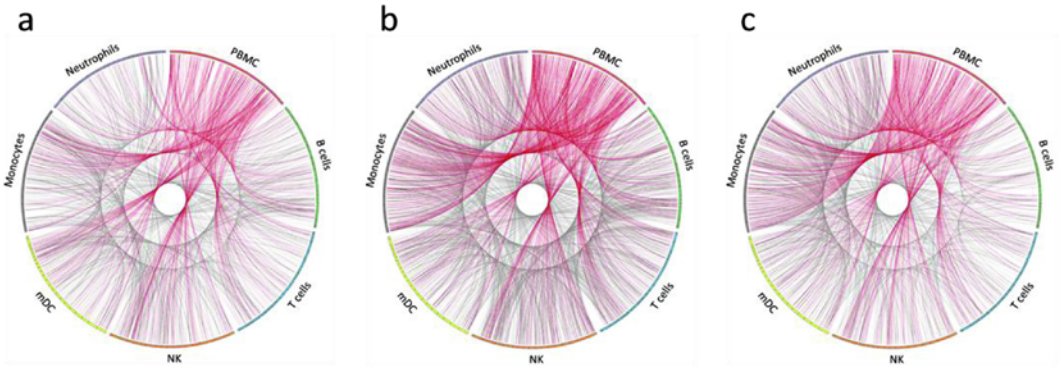



**Fig 4. Visualization of differentially expressed RNA transcripts in PBMC and individual immune cell types.** Circos plots of differentially expressed RNA transcripts from a vaccinated subject at **(a)** day 1, **(b)** day 3, and **(c)** day 7 post-TIV vaccination (fold change of $\geq$1.5x and $p \leq 0.05$). All RNA transcript classes are represented. For each cell type, the colored bar on the outer circle represents the entire human genome; segments within the bars divide the genome into chromosomes. Red lines indicate DE transcripts that are shared between PBMC and purified immune cell types. Gray lines indicate DE transcripts that are shared between the purified immune cell types.

doi:10.1371/journal.pone.0118528.g004

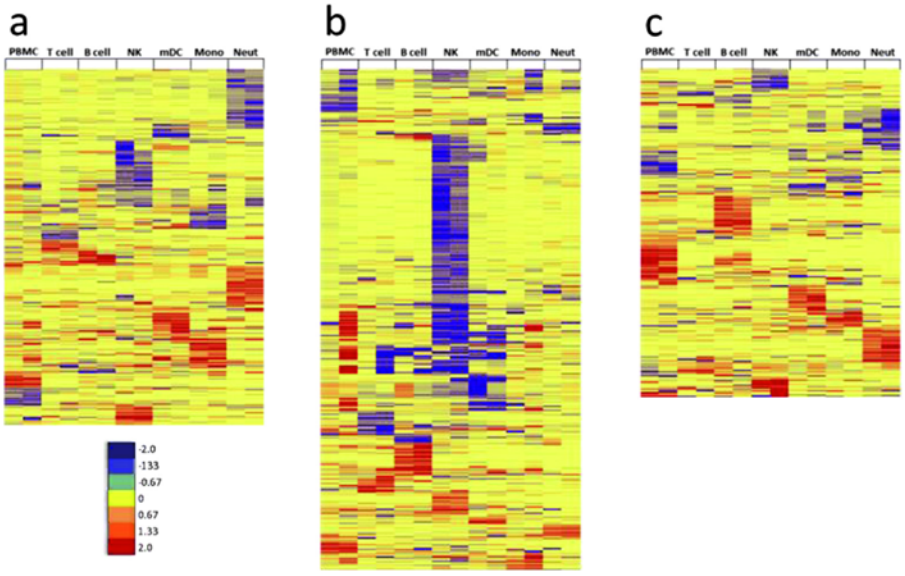



**Fig 5. Unique modules of RNA transcripts are differentially expressed in each immune cell type after TIV vaccination.** Differentially expressed RNA transcripts ($\geq$1.5-fold change, $p < 0.05$) that were shared between both subjects after TIV-vaccination were subjected to semi-supervised hierarchical clustering analysis. Log2 fold-change values of shared DE transcripts in all cell types from both subjects were clustered at **(a)** day 1 (463 transcripts), **(b)** day 3 (653 transcripts), and **(c)** day 7 (428 transcripts) post-vaccination. Very little overlap of shared differentially expressed RNA transcripts is observed between cell types; red = up; yellow = no change; blue = down.

doi:10.1371/journal.pone.0118528.g005

we only considered DE transcripts from each cell type that were shared in both subjects after TIV vaccination in further downstream investigations. Using semi-supervised hierarchical clustering, little overlap in the significantly changing protein-coding RNA transcripts was observed between each cell type and at each time point after TIV vaccination (Fig. 5A-C).

Additionally, RNA-seq analysis provided a platform to investigate differential splicing events after TIV vaccination. Using the *Multivariate Analysis of Transcript Splicing (MATS)* data analysis package[39], splicing events were identified in each cell type from each subject (S4 Table. Total splicing events identified in each cell type). Differential splicing events ($p \leq 0.05$ and FDR$\leq$0.05) were then identified in each cell type from each individual (S5 Table. Differential splicing events identified in each subject, cell type and time point). Several splicing events shared between both subjects were identified (S6 Table. Shared differential splicing events).

For proteins, the DE threshold was lowered to $\geq$1.25-fold to adjust for iTRAQ under-reporting of fold changes[44]. By choosing this threshold, we obtained comprehensive lists of DE proteins from each cell type that were shared between both subjects at each time point. Similar to RNA, there was little correlation between PBMC and purified immune cell types when comparing DE proteins (S7 Table. Comparison of differentially expressed proteins in PBMC and individual immune cell types). Circos was used to plot DE proteins from PBMC and each purified immune cell type in a vaccinated subject over the length of the human genome and to visualize overlap of differentially expressed proteins at three time points after TIV vaccination (day 1, day 3, and day 7) (Fig. 6). Similar to RNA data, the plots showed a lack of substantial overlap in DE proteins between PBMC and purified immune cell types, as well as changing patterns of overlapping expression for PBMC and each cell type at each time point after TIV vaccination. Substantial variability was observed in the number of cell type-specific DE proteins, with less than 20% being shared between both subjects for most cell types and time points (S8 Table. Shared DE proteins; S8 Dataset. Shared up-regulated DE proteins; and S9 Dataset. Shared down-regulated DE proteins). Similar to transcriptomic data, semi-supervised hierarchical clustering revealed little overlap in the shared DE proteins from each cell type at each time point after TIV vaccination (Fig. 7A-C).

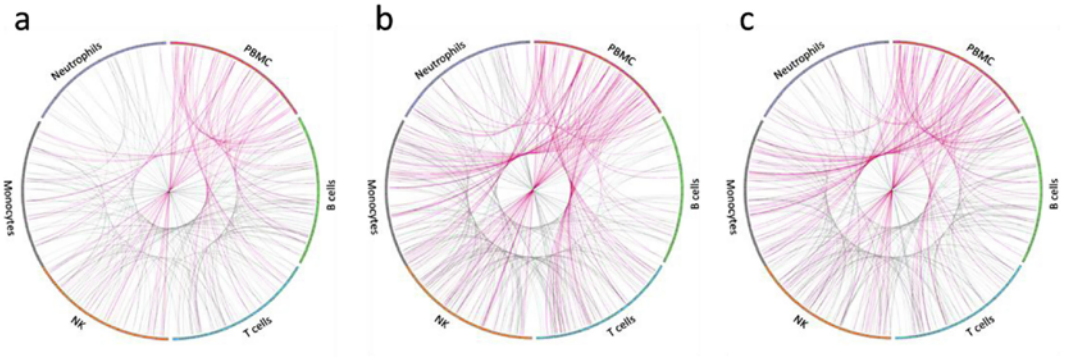



**Fig 6. Visualization of differentially expressed proteins in PBMC and individual immune cell types.** Circos plots of differentially expressed proteins from a vaccinated subject at **(a)** day 1, **(b)** day 3, and **(c)** day 7 post-TIV vaccination (fold change of $\geq$1.25x). For each cell type, the colored bar on the outer circle represents the entire human genome; segments within the bars divide the genome into chromosomes. Red lines indicate DE proteins that are shared between PBMC and purified immune cell types. Gray lines indicate DE proteins that are shared between the purified immune cell types.

doi:10.1371/journal.pone.0118528.g006

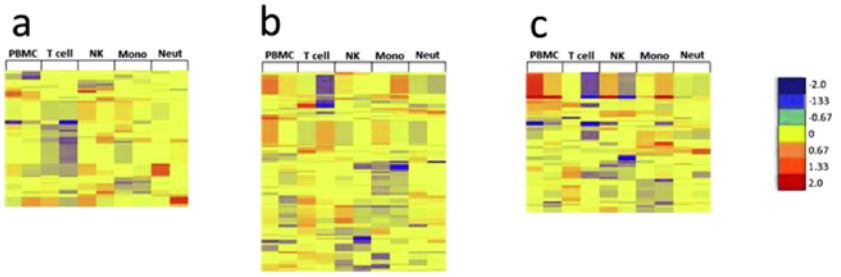



**Fig 7. Unique modules of proteins are differentially expressed in each immune cell type after TIV vaccination.** Differentially expressed proteins (≥1.25-fold change) that were shared between both subjects after vaccination with TIV were subjected to semi-supervised hierarchical clustering analysis. Log2 fold change values of shared DE proteins in each cell type from both subjects were clustered at **(a)** day 1 (196 proteins), **(b)** day 3 (263 proteins), and **(c)** day 7 (199 proteins) post-vaccination. Very little overlap of differentially expressed proteins is observed between cell types; red = up; yellow = no change; blue = down. B cell data was derived from only one subject due to insufficient recovery of B cells from the second subject.

doi:10.1371/journal.pone.0118528.g007

Following cluster analysis, lists of shared DE transcripts or proteins from each cell type and time point were loaded into *Ingenuity Pathway Analysis* (*IPA*) to identify the most significant biological interactions after TIV vaccination. When comparing the top network identified in each cell type for both protein-coding RNA transcripts and proteins (Figs. 8 and 9, respectively), each cell type induced unique biological networks at day 1 after TIV vaccination. Similarly, unique RNA and protein networks were observed in each cell type at day 3 and day 7 after vaccination (S11 Fig. Networks derived from DE RNA transcripts at d3 post-TIV vaccination; S12 Fig. Networks derived from DE RNA transcripts at d7 post-TIV vaccination; S13 Fig. Networks derived from DE proteins at d3 post-TIV vaccination; and S14 Fig. Networks derived from DE proteins at d7 post-TIV vaccination). The top biological networks and canonical pathways identified in each cell type at each time point are shown in S10 Dataset. Top networks and pathways identified in TIV-vaccinated subjects.

## Discussion

The goal of this study was to develop methods and establish protocols that can be used in future systems vaccinology studies. By utilizing this efficient cell-sorting protocol, we obtained sufficient numbers of six immune cell types purified from freshly collected whole blood to perform both RNA-sequencing and quantitative proteomics experiments. Importantly, cells were processed and stored for downstream applications in a single day, thus avoiding the pitfalls of freeze-thaw cycles on downstream analysis. In this study, sorting was stopped once the target number of cells was reached (1.5–3 $x10^6$ cells) even if MACS-enriched material remained. Collection of larger numbers of cells is therefore possible for some cell types. In this regard, we have utilized this protocol in a subsequent vaccinology study and collected up to $4x10^6$ neutrophils, up to $3x10^6$ B cells, T cells, NK and monocytes, and up to $1x10^6$ mDC from similar amounts of starting material.

Further fractionation of these six cell types into sub-populations was considered. However, we decided against this approach for several reasons. First, we were interested in broadly sampling the immune system in response to vaccination. Previous vaccinology studies that investigated responses of individual immune cells only focused on selected cell types [9,10]. Our approach profiled both transcriptomic and quantitative proteomic responses of six essential innate and adaptive immune cell types, including neutrophils and NK cells, after vaccination. Signals from small, potentially important sub-populations from any of these immune cell types

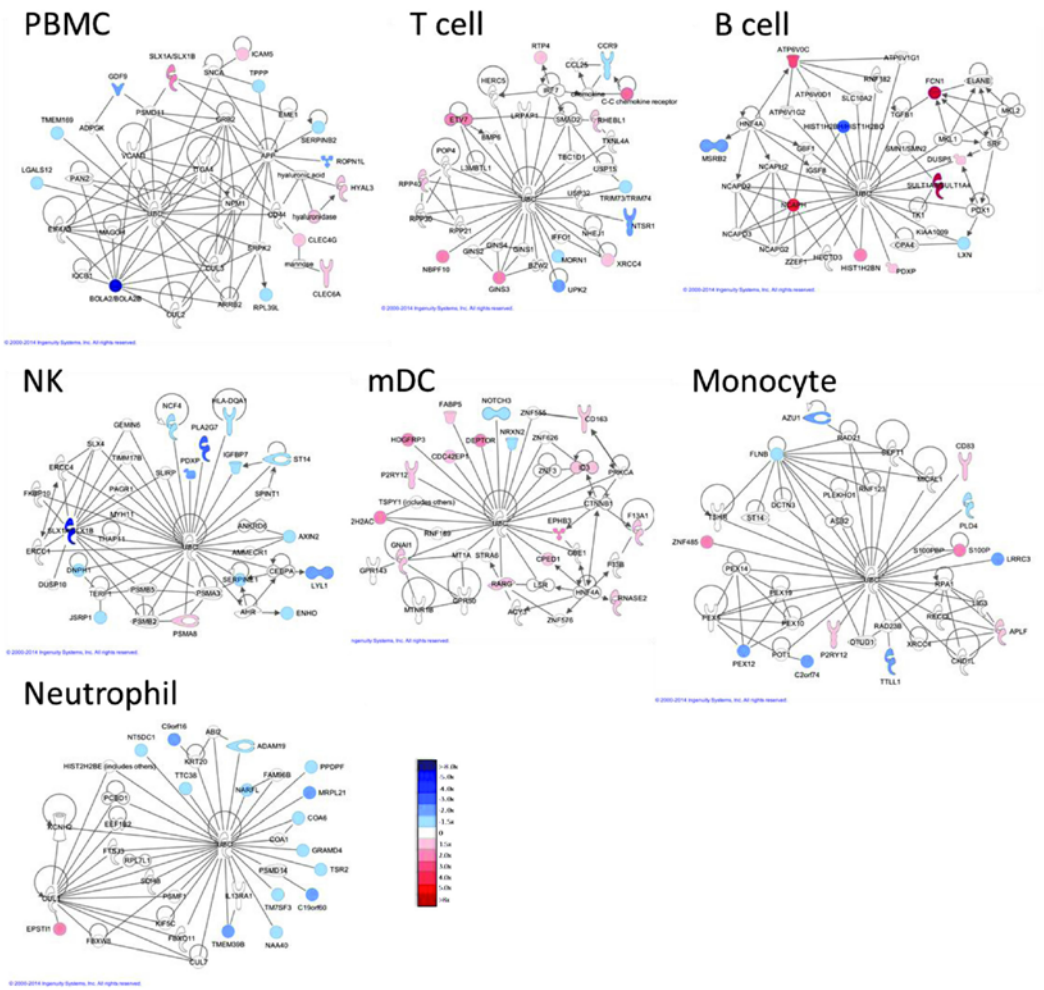



**Fig 8. Networks derived from DE RNA transcripts at d1 post-TIV vaccination.** Differentially expressed protein-coding RNA transcripts (1.5x, $p<0.05$) identified in both TIV-vaccinated subjects at day 1 post-vaccination were imported into IPA, and the top network identified in each cell type is displayed. Very little overlap of individual transcripts or biological networks that are activated is observed between cell types.

doi:10.1371/journal.pone.0118528.g008

may still be masked in our systems analysis. However, by sorting for these six immune cell types, we simultaneously investigated both innate and adaptive immune cell responses to vaccination at a cell-specific level. Second, pursuing sub-populations of immune cells would require either obtaining larger blood samples or reducing the number of distinct cell types that we could purify in order to recover sufficient cells for both RNA-Seq and proteomics analyses. If only transcriptomic studies had been performed, sorting for sub-populations from selected immune cell types would have been possible. Finally, the added cost for analysis of both

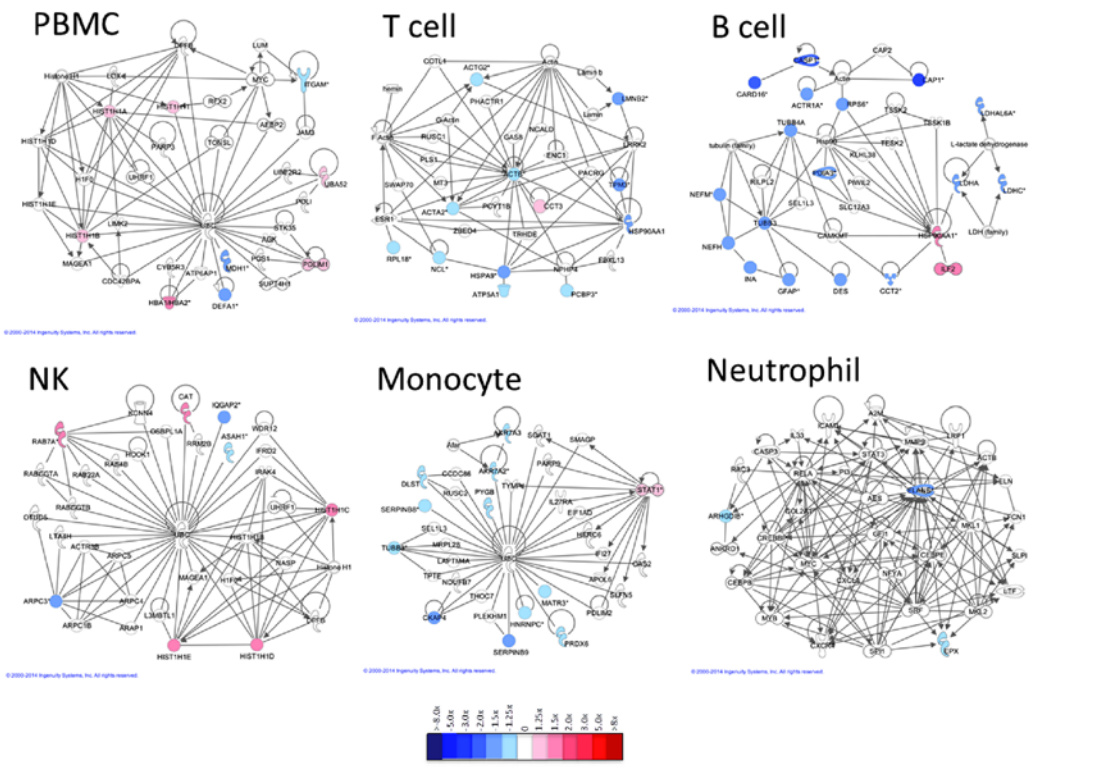



**Fig 9. Networks derived from DE proteins at d1 post-TIV vaccination.** Differentially expressed proteins (1.25x) identified in both TIV-vaccinated subjects at day 1 post-vaccination were imported into IPA, and the top network identified in each cell type is displayed (*multiple ENSPs mapped to these proteins). Very little overlap of individual proteins or biological networks that are activated is observed between cell types. B cell data was derived from only one subject due to insufficient recovery of B cells from the second subject.

doi:10.1371/journal.pone.0118528.g009

transcriptomic and proteomic data from additional sub-populations was considered prohibitive for this study's broad survey of innate and adaptive immune responses after vaccination. Future studies that focus on a specific immune cell type(s) and/or sub-populations can easily be performed by adapting our protocol, especially if only transcriptomic analysis is proposed.

Emerging technologies that allow for greater identification of sub-populations of cells and the potential for single cell analysis are now possible [45]. For example, CyTOF offers an opportunity to investigate both cell surface and intracellular protein expression at the single cell level [46]. This technology allows for staining of a potentially unlimited number of cellular markers by eliminating the spectral overlap that plagues traditional flow cytometry applications due to use of fluorescently-labeled antibodies. Therefore, analysis of a substantially increased number of cell subtypes from a single sample can be performed. However, the destructive nature of this technology (single cell ICP mass cytometry) eliminates the potential to collect live cells for further downstream applications. Additionally, by nature, antibody-targeted validation studies require that previously identified molecules be selected for screening. The approach described in this study generated both unbiased and quantitative global

transcriptomic and proteomic data from six purified immune cell populations after vaccination. CyTOF offers a powerful, single-cell, high throughput approach to validate and characterize results derived by these types of systems studies.

This study optimized our strategy to generate and analyze RNA-seq and quantitative proteomics data from individual immune cell types sorted from fresh human blood. Differential analysis for each immune cell type revealed unique transcriptomic and proteomic expression profiles as well as changing biological networks during the early response after vaccination. Lending support to our strategy, previous transcriptional findings from systems analysis after TIV vaccination were identified in our approach. For example, we found that B cell-specific transcripts identified by Nakaya *et al.* as correlative predictors of protective immunity following TIV vaccination, including immunoglobulin genes and TNFSFR17 as well as the transcription factor XBP-1, were up-regulated in sorted B cell samples from both of our subjects 7 days after vaccination (S1 Dataset)[9]. Additionally, we found that CXCR3, the receptor for CXCL10/IP-10, was significantly up-regulated in both PBMC and sorted B cell samples after TIV vaccination (S1 Dataset); CXCL10/IP-10 was the only cytokine Nakaya *et al.* identified as being significantly increased in the serum of TIV-vaccinated subjects in their systems study[9]. These data suggest that our subject cohort likely attained at least some measure of protection after TIV vaccination. Future studies using these protocols will correlate vaccine-induced differential expression of both RNA and proteins, as well as serum cytokine levels, with day 28 antibody titers to make predictions about generation of protective immunity in response to vaccination.

The methods and strategies developed in this project provided a unique and important opportunity to investigate the quantitative and qualitative differences between PBMC and individual immune cell types at both the transcriptomic and proteomic levels. By utilizing RNA-seq rather than microarray analysis, we were able to identify and quantify an expanded fraction of the transcriptome, which included 29 different classes of RNA transcripts. Additionally, both transcriptomic and proteomic data were visualized across the human reference genome sequence. Only a small fraction of differentially expressed transcripts and proteins identified in the purified immune cell types were also identified in the PBMC fraction. Thus, by analyzing each cell type individually, cell-specific transcriptomic and proteomic contributions to the immune response following vaccination were identified. This cell type-specific information, coupled with unbiased systems biology approaches, provides a more comprehensive approach to monitor and eventually model vaccine responses. The approaches developed in this pilot project will help to guide future systems biology studies aimed at modeling and predicting complex responses to vaccines and vaccine adjuvants involving interactions between multiple cell types.

## Supporting Information

**S1 Dataset. RNA-seq quality control.**
(XLSX)

**S2 Dataset. Normalized transcript expression in human immune cells prior to and post-TIV vaccination.**
(XLSX)

**S3 Dataset. Normalized transcript expression in human immune cells filtered for an RPKM of 1.0 in at least one sample from one subject.**
(XLSX)

**S4 Dataset. Normalized protein expression in human immune cells prior to and post-TIV vaccination.**
(XLSX)

**S5 Dataset. Normalized protein expression in human immune cells filtered to remove zero values and contaminating keratins from one subject.**
(XLSX)

**S6 Dataset. Shared up-regulated DE RNA transcripts.**
(XLSX)

**S7 Dataset. Shared down-regulated DE RNA transcripts.**
(XLSX)

**S8 Dataset. Shared up-regulated DE proteins.**
(XLSX)

**S9 Dataset. Shared down-regulated DE proteins.**
(XLSX)

**S10 Dataset. Top networks and pathways identified in TIV-vaccinated subjects.**
(XLSX)

**S1 Fig. RNA quality control.** Scatter plots showing the correlation of total RNA transcripts between time points and subjects. **(a)** Time point comparison within the same subject (HD30 PBMC day 3 vs HD30 PBMC day 0). **(b)** Subject-to-subject comparison of one time point (HD30 PBMC day 3 vs HD31 PBMC day 3). Both comparisons show correlation greater than 0.95.
(TIF)

**S2 Fig. Proteomics quality control. (a)** Scatter plot showing the protein abundances measured in two technical replicates of the ICCS common control. Each dot represents an individual protein. X axis represents the protein abundance measured in replicate 2. Y-axis represents the protein abundances measured in replicate 1. **(b)** Scatter plot showing the distribution of fold changes of proteins with respect to their abundances. Each dot represents an individual protein. X axis represents protein abundance. Y axis represents fold changes. **(c)** Cluster dot plot showing the distribution of fold changes in different iTRAQ channels. Each dot represents an individual protein and the lines represent patterns of expression change.
(TIF)

**S3 Fig. Flow chart for immune cell purification. (a)** When 150–300x$10^6$ PBMC were obtained, B cells ($CD19^+$), monocytes ($CD14^+$) and T cells ($CD3^+$) were first positively selected from the PBMC fraction by MACS; approximately 15% of PBMC were dedicated for $CD3^+$ enrichment, 35% of PBMC were dedicated to $CD14^+$ enrichment, and 45% of PBMC were dedicated to $CD19^+$ enrichment. Negative flow through material was collected, pooled and subsequently depleted of remaining $CD3^+$, $CD14^+$, $CD15^+$, and $CD19^+$ cells to enrich for mDC and NK cells. All MACS enriched cell populations were stained as in Fig. 1A with the addition of 7-AAD for live/dead cell identification and subjected to FACS sorting to yield highly purified cell populations. **(b)** When >300x$10^6$ PBMC were obtained, CD3+, CD19+ and CD14+ selection was performed as in (a), with a smaller cell fraction dedicated to each sort, while NK and mDC were enriched by negative selection directly from PBMC. Cells were stained and FACS sorted as in (a). **(c)** When <150x$10^6$ PBMC were obtained, all PBMC were dedicated to $CD19^+$ B cell selection. The CD19-negative flow through was then subjected to $CD3^+CD14^+$ dual

positive selection. MACS enriched cells were stained as in (a), and B cells were FACS sorted from the $CD19^+$ fraction, T cells and monocytes were FACS sorted from the $CD3^+CD14^+$ fraction, and NK and mDC were FACS sorted from the $CD19^-CD3^-CD14^-$ fraction. Any potential contaminating neutrophils were eliminated from the NK and mDC fraction by staining with anti-CD15 during FACS sorting.
(TIF)

**S4 Fig. Individual cell types are not activated by the sorting process.** Aliquots of whole blood (WB), PBMC and pooled sorted cells ($\sim$10,000 each cell type) from a representative subject were stained with antibodies directed against CD3, CD11c, CD14, CD15, CD19 and CD56 for phenotyping as in Fig. 1A, as well as CD69, CD86 and CD134 to measure cellular activation. Fluorescence minus one (FMO) controls were used to determine background fluorescence levels for activation marker staining in each cell type from WB and PBMC samples. Assessment of surface expression (mean fluorescence intensity; MFI) of **(a)** CD69 in each cell type, **(b)** CD86 in monocyes, B cells, and mDC, and **(c)** CD134 in T cells reveals that none of the cell types were significantly activated during any step of our sorting protocol.
(TIF)

**S5 Fig. Adequate RNA quantity and quality is obtained from sorted immune cells for RNA-seq applications.** RNA isolated from sorted immune cells (500,000 each cell type except mDC, which contained 400,000 at d0, 567,000 at d1, 438,000 at d3, and 548,000 at d7) from a single vaccinated subject was quantified (top panel) and evaluated for RNA integrity (bottom panel) as described in Materials and Methods.
(TIF)

**S6 Fig. Transcriptional profiling of PBMC and individual immune cell types.** Baseline, day 0 RNA profiles of PBMC and each purified cell type (all transcript classes represented, non-zero transcripts with an RPKM of 1 in at least one sample; $\sim$21,000 transcripts) from a single subject were plotted using Circos to visualize relative expression of transcripts across the genome. Bars on the outside of the circle represent individual chromosomes. The heat-map color scaling parameter was set to "scale_log_base = 1" to allow for optimal color space.
(TIF)

**S7 Fig. Adequate protein quantity is obtained from sorted immune cells for proteomics applications.** Total protein isolated from sorted immune cells ($1x10^6$ each cell type) from a single vaccinated subject was quantified as described in Materials and Methods.
(TIF)

**S8 Fig. Two iTRAQ strategies for quantitative proteomic analysis of immune cells after vaccination. (a)** Experimental design. In strategy 1, multiple immune cell types from one time point were multiplexed together in the experiment. In strategy 2, different time points from the same immune cell type were multiplexed together. An immune cell common standard (ICCS) was used to normalize reporter ion intensities across the experiments. **(b)** Unsupervised hierarchical clustering analysis and **(c)** PCA of pseudo-spectral counts from one subject generated using strategy 1 (left panels; 5,676 proteins, filtered to remove zero values and contaminating keratins) or strategy 2 (right panels, 3,852 proteins, filtered to remove zero values and contaminating keratins) reveals that cell-types cluster together and display distinct cell-type specific patterns of protein expression using strategy 2, but not with strategy 1.
(TIF)

**S9 Fig. Proteomic profiling of PBMC and individual immune cell types.** Baseline, day 0 protein profiles of PBMC and each purified cell type (3,852 proteins) from a single subject were

plotted using Circos to visualize relative expression of proteins across the genome. Bars on the outside of the circle represent individual chromosomes. The heat-map color scaling parameter was set to "scale_log_base = 10" to allow for optimal color space.
(TIF)

**S10 Fig. Principal component analysis reveals poor correlation of proteomes between subjects. (a)** RNA transcripts (all RNA classes represented, filtered to remove zero values; ~39,106 total transcripts) and **(b)** proteins (5,304 total proteins, filtered to remove zero values and contaminating keratins) from subject 1 (HD31; large circles) and subject 2 (HD30; small circles) were clustered in the same experiment. RNA from both subjects clusters similarly, while proteins do not.
(TIF)

**S11 Fig. Networks derived from DE RNA transcripts at d3 post-TIV vaccination.** Differentially expressed protein-coding RNA transcripts (1.5x, $p<0.05$) identified in both TIV-vaccinated subjects at day 3 post-vaccination were imported into *IPA*, and the top network identified in each cell type is displayed. Very little overlap of individual transcripts or biological networks that are activated is observed between cell types.
(TIF)

**S12 Fig. Networks derived from DE RNA transcripts at d7 post-TIV vaccination.** Differentially expressed protein-coding RNA transcripts (1.5x, $p<0.05$) identified in both TIV-vaccinated subjects at day 7 post-vaccination were imported into *IPA*, and the top network identified in each cell type is displayed. Very little overlap of individual transcripts or biological networks that are activated is observed between cell types.
(TIF)

**S13 Fig. Networks derived from DE proteins at d3 post-TIV vaccination.** Differentially expressed proteins (1.25x) identified in both TIV-vaccinated subjects at day 3 post-vaccination were imported into *IPA*, and the top network identified in each cell type is displayed (∗multiple ENSPs mapped to these proteins). Very little overlap of individual proteins or biological networks that are activated is observed between cell types. B cell data was derived from only one subject due to insufficient recovery of B cells from the second subject.
(TIF)

**S14 Fig. Networks derived from DE proteins at d7 post-TIV vaccination.** Differentially expressed proteins (1.25x) identified in both TIV-vaccinated donors at day 7 post-vaccination were imported into *IPA*, and the top network identified in each cell type is displayed (∗multiple ENSPs mapped to these proteins). Very little overlap of individual proteins or biological networks that are activated is observed between cell types. B cell data was derived from only one subject due to insufficient recovery of B cells from the second subject.
(TIF)

**S1 Table. Summary of baseline RNA transcripts identified in each cell type from one subject by RNA-seq analysis.**
(TIF)

**S2 Table. Comparison of differentially expressed RNA transcripts in PMBC and individual immune cell types.**
(TIF)

**S3 Table. Shared DE RNA transcripts (all transcript classes represented).**
(TIF)

**S4 Table. Total splicing events identified in each cell type.**
(TIF)

**S5 Table. Differential splicing events identified in each subject, cell type and time point.**
(TIF)

**S6 Table. Shared differential splicing.**
(TIF)

**S7 Table. Comparison of differentially expressed proteins in PMBC and individual immune cell types.**
(TIF)

**S8 Table. Shared DE proteins.**
(TIF)

## Acknowledgments

This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, Tennessee. Flow Cytometry experiments were performed in the VMC Flow Cytometry Shared Resource. The VMC Flow Cytometry Shared Resource is supported by the Vanderbilt Ingram Cancer Center (P30 CA68485) and the Vanderbilt Digestive Disease Research Center (DK058404). We thank Attila Csordas and the PRIDE team for assistance uploading the proteomics data sets to the ProteomXchange consortium PRIDE database.

## Author Contributions

Conceived and designed the experiments: KLH PS SJ KME AJL. Performed the experiments: KLH PS NP TMA KAF. Analyzed the data: KLH PS XN NP AG QL YG. Contributed reagents/materials/analysis tools: YG YS SL. Wrote the paper: KLH PS AG AJL. Wrote the parental study protocol, obtained IRB approval, collected blood samples from subjects and processed blood samples: LMH.

## References

1. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999; 402: C47–52. PMID: 10591225
2. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet. 2001; 2: 343–372. PMID: 11701654
3. Kitano H. Systems biology: Toward system-level understanding of biological systems. In: Kitano H, editor. Foundations of systems biology Edition 1. Cambridge, Massachusetts MIT Press; 2001. pp. 1–36.
4. Pulendran B, Li S, Nakaya HI. Systems vaccinology. Immunity. 2010; 33: 516–529. doi: 10.1016/j.immuni.2010.10.006 PMID: 21029962
5. Trautmann L, Sekaly RP. Solving vaccine mysteries: a systems biology perspective. Nat Immunol. 2011; 12: 729–731. doi: 10.1038/ni.2078 PMID: 21772284
6. Gaucher D, Therrien R, Kettaf N, Angermann BR, Boucher G, Filali-Mouhim A, et al. Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. J Exp Med. 2008; 205: 3119–3131. doi: 10.1084/jem.20082292 PMID: 19047440
7. Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol. 2009; 10: 116–125. doi: 10.1038/ni.1688 PMID: 19029902
8. Bucasas KL, Franco LM, Shaw CA, Bray MS, Wells JM, Nino D, et al. Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans. J Infect Dis. 2011; 203: 921–929. doi: 10.1093/infdis/jiq156 PMID: 21357945

9. Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, et al. Systems biology of vaccination for seasonal influenza in humans. Nat Immunol. 2011; 12: 786–795. doi: 10.1038/ni.2067 PMID: 21743478

10. Obermoser G, Presnell S, Domico K, Xu H, Wang Y, Anguiano E, et al. Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. Immunity. 2013; 38: 831–844. doi: 10.1016/j.immuni.2012.12.008 PMID: 23601689

11. Tan Y, Tamayo P, Nakaya H, Pulendran B, Mesirov JP, Haining WN. Gene signatures related to B-cell proliferation predict influenza vaccine-induced antibody response. Eur J Immunol. 2014; 44: 285–295. doi: 10.1002/eji.201343657 PMID: 24136404

12. Tsang JS, Schwartzberg PL, Kotliarov Y, Biancotto A, Xie Z, Germain RN, et al. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. Cell. 2014; 157: 499–513. doi: 10.1016/j.cell.2014.03.031 PMID: 24725414

13. Furman D, Jojic V, Kidd B, Shen-Orr S, Price J, Jarrell J, et al. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. Mol Syst Biol. 2013; 9: 659. doi: 10.1038/msb.2013.15 PMID: 23591775

14. Banchereau R, Jordan-Villegas A, Ardura M, Mejias A, Baldwin N, Xu H, et al. Host immune transcriptional profiles reflect the variability in clinical disease manifestations in patients with Staphylococcus aureus infections. PLoS One. 2012; 7: e34390. doi: 10.1371/journal.pone.0034390 PMID: 22496797

15. Berry MP, Graham CM, McNab FW, Xu Z, Bloch SA, Oni T, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature. 2010; 466: 973–977. doi: 10.1038/nature09247 PMID: 20725040

16. Bloom CI, Graham CM, Berry MP, Wilkinson KA, Oni T, Rozakeas F, et al. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. PLoS One. 2012; 7: e46191. doi: 10.1371/journal.pone.0046191 PMID: 23056259

17. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, et al. Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood. 2007; 109: 2066–2077. PMID: 17105821

18. Tattermusch S, Skinner JA, Chaussabel D, Banchereau J, Berry MP, McNab FW, et al. Systems biology approaches reveal a specific interferon-inducible signature in HTLV-1 associated myelopathy. PLoS Pathog. 2012; 8: e1002480. doi: 10.1371/journal.ppat.1002480 PMID: 22291590

19. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell. 5th edition. New York, NY: Garland Science; 2007.

20. Reimold AM, Iwakoshi NN, Manis J, Vallabhajosyula P, Szomolanyi-Tsuda E, Gravallese EM, et al. Plasma cell differentiation requires the transcription factor XBP-1. Nature. 2001; 412: 300–307. PMID: 11460154

21. Pulendran B, Ahmed R. Immunological mechanisms of vaccination. Nat Immunol. 2011; 12: 509–517. PMID: 21739679

22. Sallusto F, Lanzavecchia A, Araki K, Ahmed R. From vaccines to memory and back. Immunity. 2010; 33: 451–463. doi: 10.1016/j.immuni.2010.10.008 PMID: 21029957

23. Debey-Pascher S, Hofmann A, Kreusch F, Schuler G, Schuler-Thurner B, Schultze JL, et al. RNA-stabilized whole blood samples but not peripheral blood mononuclear cells can be stored for prolonged time periods prior to transcriptome analysis. J Mol Diagn. 2011; 13: 452–460. doi: 10.1016/j.jmoldx.2011.03.006 PMID: 21704280

24. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10: R25. doi: 10.1186/gb-2009-10-3-r25 PMID: 19261174

25. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25: 1105–1111. doi: 10.1093/bioinformatics/btp120 PMID: 19289445

26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

27. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11: R25. doi: 10.1186/gb-2010-11-3-r25 PMID: 20196867

28. Wang H, Qian WJ, Mottaz HM, Clauss TR, Anderson DJ, Moore RJ, et al. Development and evaluation of a micro- and nanoscale proteomic sample preparation method. J Proteome Res. 2005; 4: 2397–2403. PMID: 16335993

29. Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, et al. Measurement of protein using bicinchoninic acid. Anal Biochem. 1985; 150: 76–85. PMID: 3843705

30. Browne CM, Samir P, Fites JS, Villarreal SA, Link AJ. The yeast eukaryotic translation initiation factor 2B translation initiation complex interacts with the fatty acid synthesis enzyme YBR159W and

endoplasmic reticulum membranes. Mol Cell Biol. 2013; 33: 1041–1056. doi: 10.1128/MCB.00811-12 PMID: 23263984

31. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, et al. Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol. 1999; 17: 676–682. PMID: 10404161
32. Eng JK, Fischer B, Grossmann J, Maccoss MJ. A fast SEQUEST cross correlation algorithm. J Proteome Res. 2008; 7: 4598–4602. doi: 10.1021/pr800420s PMID: 18774840
33. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994; 5: 976–989. doi: 10.1016/1044-0305(94)80016-2 PMID: 24226387
34. R Core Team. R: A Language and Environment for Statistical Computing. 2014; http://www.R-project.org
35. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95: 14863–14868. PMID: 9843981
36. Saldanha AJ. Java Treeview—extensible visualization of microarray data. Bioinformatics. 2004; 20: 3246–3248. PMID: 15180930
37. Strand. Avadis NGS v1.5 Reference Manual. Strand. 2013; http://www.strand-ngs.com/sites/default/files/private/manual/AvadisNGS-Reference-manual-v1.5.pdf
38. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological. 1995; 57: 289–300.
39. Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q, et al. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res. 2012; 40: e61. doi: 10.1093/nar/gkr1291 PMID: 22266656
40. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19: 1639–1645. doi: 10.1101/gr.092759.109 PMID: 19541911
41. Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database (Oxford). 2011; 2011: bar049. doi: 10.1093/database/bar049 PMID: 22083790
42. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18: 1509–1517. doi: 10.1101/gr.079558.108 PMID: 18550803
43. Brown SM, Goecks J, Taylor J. RNA sequencing with NGS. In: Brown SM, editor. Next-generation DNA sequencing informatics. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2013. pp. 171–186.
44. Ow SY, Salim M, Noirel J, Evans C, Rehman I, Wright PC. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". J Proteome Res. 2009; 8: 5347–5355. doi: 10.1021/pr900634c PMID: 19754192
45. Mullard A. Single-cell profiling sheds new light. Nat Rev Drug Discov. 2011; 10: 477–478. doi: 10.1038/nrd3487 PMID: 21720392
46. Cheung RK, Utz PJ. Screening: CyTOF-the next generation of cell detection. Nat Rev Rheumatol. 2011; 7: 502–503 doi: 10.1038/nrrheum.2011.110 PMID: 21788983

# Appendix Y – Manuscript – 4: Sculpting MHC class II-restricted self and non-self peptidome by the class I Ag-processing machinery and its impact on Th-cell responses

 DOI: 10.1002/eji.201243087 Eur. J. Immunol. 2013. 43: 1162–1172

European Journal of Immunology

# Sculpting MHC class II–restricted self and non-self peptidome by the class I Ag-processing machinery and its impact on Th-cell responses

*Charles T. Spencer[1], Srdjan M. Dragovic[1], Stephanie B. Conant[1], Jennifer J. Gray[1], Mu Zheng[2], Parimal Samir[2], Xinnan Niu[2], Magdalini Moutaftsi[3], Luc Van Kaer[1], Alessandro Sette[3], Andrew J. Link[1,2] and Sebastian Joyce[1]*

[1] Department of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, TN, USA
[2] Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN, USA
[3] Center for Infectious Diseases, Allergy and Asthma Research, La Jolla Institute of Allergy and Immunology, La Jolla, CA, USA

It is generally assumed that the MHC class I antigen (Ag)-processing (CAP) machinery — which supplies peptides for presentation by class I molecules — plays no role in class II–restricted presentation of cytoplasmic Ags. In striking contrast to this assumption, we previously reported that proteasome inhibition, TAP deficiency or ERAAP deficiency led to dramatically altered T helper (Th)-cell responses to allograft (HY) and microbial (*Listeria monocytogenes*) Ags. Herein, we tested whether altered Ag processing and presentation, altered $CD4^+$ T-cell repertoire, or both underlay the above finding. We found that TAP deficiency and ERAAP deficiency dramatically altered the quality of class II-associated self peptides suggesting that the CAP machinery impacts class II–restricted Ag processing and presentation. Consistent with altered self peptidomes, the $CD4^+$ T-cell receptor repertoire of mice deficient in the CAP machinery substantially differed from that of WT animals resulting in altered $CD4^+$ T-cell Ag recognition patterns. These data suggest that TAP and ERAAP sculpt the class II–restricted peptidome, impacting the $CD4^+$ T-cell repertoire, and ultimately altering Th-cell responses. Together with our previous findings, these data suggest multiple CAP machinery components sequester or degrade MHC class II–restricted epitopes that would otherwise be capable of eliciting functional Th-cell responses.

**Keywords:** Antigen presentation · Mass spectrometry · MHC · Self peptidome · T helper (Th) cells

Additional supporting information may be found in the online version of this article at the publisher's web-site

Correspondence: Dr. Charles T. Spencer
e-mail: ctspencer@utep.edu

## Introduction

$CD4^+$ T helper (Th) cells regulate multiple cellular and humoral responses to pathogenic microbes and parasites to protect against infectious diseases. These cells sense infections by recognizing short microbial peptides presented by MHC class II molecules on the cell surface of APCs. Hence, alterations or deficiencies in factors that control class II–restricted Ag processing and presentation can alter the display of self and microbial peptides by APCs. Alterations in the presented self peptide repertoire (peptidome) can change the $CD4^+$ T-cell repertoire activated in response to an infection, which in turn can affect the host's susceptibility to infectious disease.

Th cells recognize endogenous cytosolic as well as exogenous Ags. The mechanisms controlling exogenous class II–restricted Ag presentation are quite well established [1, 2]. Nonetheless, endogenous cytosolic Ag presentation by class II molecules is less well understood. Endogenous cytosolic Ags existing within professional APCs are presented by class II molecules when they are delivered to the endo/lysosomes. These Ags are delivered to these compartments by various autophagic mechanisms — macroautophagy [3–7] or chaperone-mediated autophagy [8–10] — and processed therein for presentation to $CD4^+$ T cells [11–17]. Alternatively, cytosolic Ags expressed by class II-negative cells — such as allograft, tumour and infected cells — are acquired by phagocytosis. Professional class II-positive APCs (e.g. DCs and macrophages (MΦs)) phagocytose dying cells and process Ags into short peptides within the phago-lysosomes, assemble with class II molecules and are displayed at the cell surface [18–20]. This process, termed indirect presentation, was originally described to explain solid organ allograft rejection.

Newer data suggest that this dogmatic separation of class I and class II Ag processing and presentation is not so absolute. Interdependence between these two processing pathways has been observed either within the presenting APCs or in damaged neighbouring (donor) cells. As we reported previously, class II–restricted cytosolic Ags are exposed to modification by components of the MHC class I Ag processing (CAP) machinery in both the presenting and donor cells [21]. This modification is evident in animal models deficient in the CAP components TAP and ERAAP where an altered basal class I–restricted peptide repertoire is displayed [22–26]. However, the effect of their absence on the class II–restricted peptide repertoire has not been fully explored. Certain class II–restricted Ags, including several self peptides, that are dependent upon the actions of the CAP machinery have been identified [12–15, 21, 27–31]. Nonetheless, other investigators have not seen a dependence upon this processing machinery for class II–restricted Ag presentation [17, 32–34]. Despite the identification of a few peptides that depend on CAP machinery for presentation, the global impact the CAP machinery has on the self and nonself peptidome remains unknown. Moreover, although previous studies have observed differences in Ag presentation, no notable alterations in the frequencies of TCR Vβ usage in TAP-deficient animals for either $CD4^+$ or $CD8^+$ T cells were observed [35]. It is therefore unclear whether the class II–restricted $CD4^+$ T-cell repertoire is impacted by the CAP machinery.

We recently showed that $CD4^+$ T cell recognition of indirectly presented cytosolic, class II–restricted self (HY minor histocompatibility Ag) and non-self (*Listeria monocytogenes* (*Lm*)) peptides was enhanced in the absence of the CAP components TAP and ERAAP [21]. Curiously however, the donated HY alloantigen entered the cytosol of acceptor APCs and required LMP2-dependent immunoproteasomes for presentation [21]. Moreover, the effects of CAP components on HY alloantigen presentation were neither due to competition between class I and class II Ags nor due to competition between $CD4^+$ and $CD8^+$ T cells. They were also not caused by enhanced MHC class II, B7.1, B7.2, calreticulin or HSP90 expression nor enhanced macro-autophagy, or enhanced ER-associated degradation. Hence, we concluded from that study that the CAP machinery must regulate the quantity and/or quality of peptides available for presentation by class II molecules. Hence, we hypothesized that by regulating the class II–restricted peptidome, CAP components could alter the robustness of the Th-cell response to class II–restricted Ags [21].

We now report direct evidence that TAP and ERAAP influence the available class II-associated peptide pool. In their absence, a nearly unique self peptidome is displayed by $H2A^b$ molecules. These findings emerged from amino acid sequence analyses of the class II–associated self peptidomes isolated from WT, $TAP^{-/-}$, or $ERAAP^{-/-}$ splenocytes. As previously described [35], we also found insubstantial alterations in the TCR Vβ usage. Nonetheless, we observed significant changes within the Ag-binding CDR3 of TCR β-chains (CDR3β) expressed by $CD4^+$ T cells. Consistent with altered Ag processing and presentation and an altered TCR diversity, we found that functional Th-cell responses to $H2A^b$-restricted vaccinia viral (VACV) epitopes were also altered. $TAP^{-/-}$ mice recognized novel epitopes not recognized by WT mice and, conversely, had lost recognition of some epitopes recognized by WT mice. Our in-depth analysis of the self peptidome, mature TCR repertoire and Th-cell responses suggests that the CAP machinery meaningfully sculpts class II–restricted Ag presentation likely through sequestration or degradation of potential epitopes.

## Results

### TAP and ERAAP sculpt the class II-restricted self peptidome

Previous reports have documented an altered endogenous class I–associated self class II-restricted peptidome in the absence of the CAP components TAP or ERAAP [22–26]. Recently, increasing interdependence of the class I– and class II–restricted Ag-processing pathways and the identification of several class II–restricted peptides that require the activity of components of the CAP machinery have been reported [12–15, 27–31]. This led us to query whether the basal class II–associated self peptidome might also have a similar dependence on TAP and/or ERAAP. To this end, class II–associated peptides were eluted from affinity purified $H2A^b$ molecules expressed by WT,

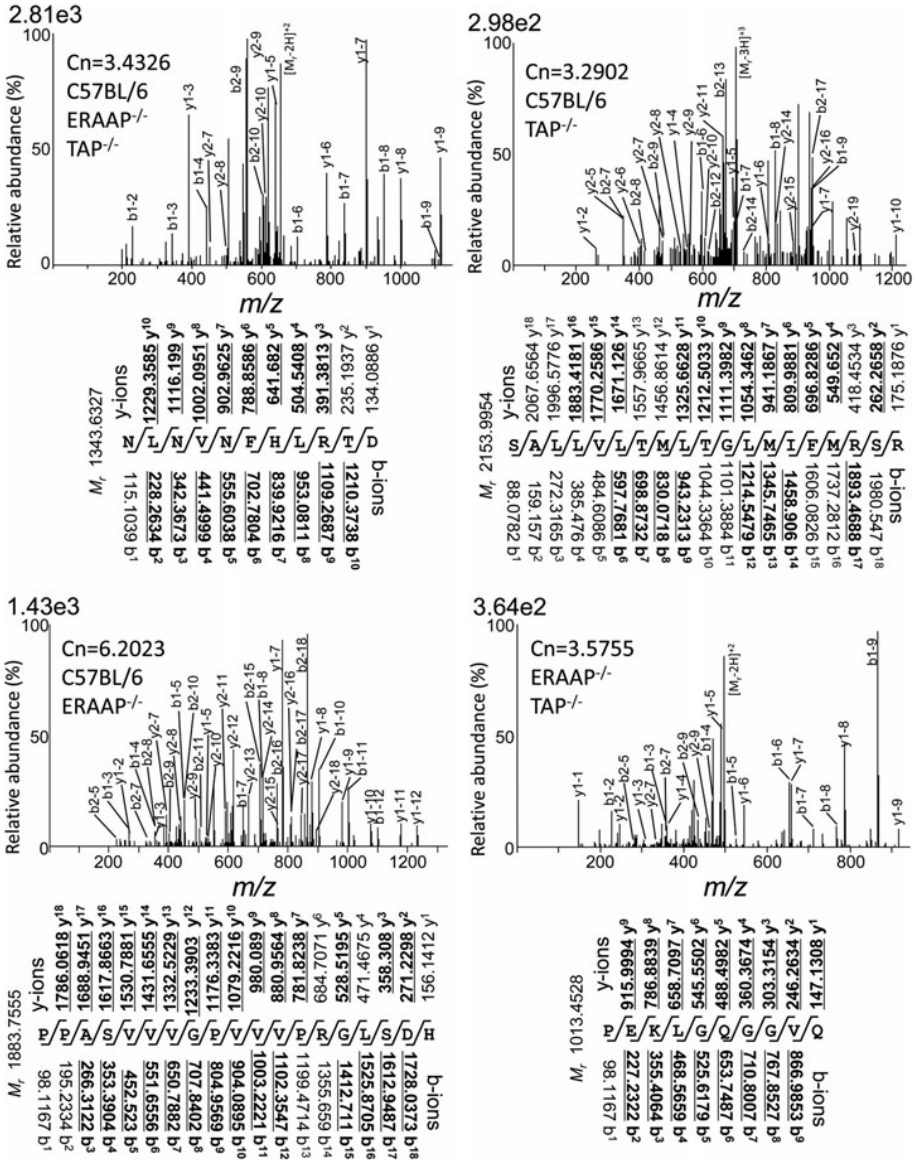



**Figure 1.** LC-MS/MS spectra of $H2A^b$-associated peptides commonly displayed by WT, $TAP^{-/-}$ and $ERAAP^{-/-}$ splenocytes. Peptides eluted from immunoaffinity purified $H2A^b$ molecules expressed by splenocytes from 68 to 70 WT, $TAP^{-/-}$ and $ERAAP^{-/-}$ mice were separated by RPC and their amino acid sequence determined by LC-MS/MS. Representative mass spectra are presented. For each spectrum, the *b*- and y-ions are indicated along with the Sequest cross-correlation score (*Cn*) showing the degree of concordance between the observed and expected fragment ions. Within the spectrum, b1, b2, y1 and y2 refer to fragment ions that have mass/charge (*m/z*) +1 or +2. Below each spectrum are the +1 ion *m/z* values for each peptide (bold underlined, observed ion masses). Note: the +2 ion mass/charge values are provided in Supporting Information Fig. 1.

B6.129-TAP$^{-/-}$, B6.129-ERAAP$^{-/-}$, and B6.129-H2Ab$^{-/-}$ splenocytes. Importantly, deficiency in either TAP or ERAAP did not alter the frequency of APCs within the spleen. Nor was the cell surface phenotype (e.g. class II and co-receptor CD80 and CD86 expression) different than WT (data not shown; [24, 25]). The recovered peptides were fractionated by reversed-phase chromatography (RPC) and their sequence deduced by LC-MS/MS tandem mass spectrometry (Fig. 1 and Supporting Information Fig. 1).

The mass/charge ($m/z$) pattern generated by MS/MS was compared against a dataset consisting of the $m/z$ patterns of theoretical and known peptide sequences. The degree of concordance between these two patterns was assigned a cross correlation score $X_{corr}$ ($Cn$). Higher $Cn$ values are assigned to those peptides whose $m/z$ pattern showed greater concordance between the observed and expected $m/z$ patterns [36]. Only peptides with a $Cn$ > 1.5 were considered to be possible peptide sequences. However, the larger the $Cn$ value the more confidence is placed in the peptide sequence identification. In addition, greater differences in the $Cn$ values between the top two most likely peptide sequence identifications ($\Delta Cn$) provides greater confidence in the identification. Therefore, peptides with a highly confident identification were considered to have a $Cn$ score >3.0 and $\Delta Cn$ >0.2. Overall, this dataset had an average $Cn$ = 3.536 and $\Delta Cn$ = 0.324. In addition, 44% of the peptides had only a single possible sequence identification for which no $\Delta Cn$ can be calculated.

To ascertain the specificity of the bound peptides, materials eluted from control H2A$^b$-deficient cells were isolated and analysed by the same methods. We found that only ~7% of the peptide sequences ($Cn$ > 1.5) identified in WT, TAP$^{-/-}$ and ERAAP$^{-/-}$ samples were also present in the control H2Ab$^{-/-}$ eluates (data not shown). These were largely derived from three sources; (i) Ig – likely representing the Ab used for immunoaffinity purification or splenic Ig that bound to protein A Sepharose used to prepare the immunoaffinity column; (ii) complement – perhaps because they bind Ig; and (iii) fibronectin, fibrinogen and other secreted proteins – likely representing unspecific contaminants of the purification. Few peptides were derived from cytosolic/intracellular proteins. Hence, peptide sequences that matched those isolated from H2Ab$^{-/-}$ splenocytes were considered an artefact of the purification. Such peptide sequences with $Cn$ > 1.5 when present in WT, TAP$^{-/-}$ and ERAAP$^{-/-}$ samples were removed from all downstream analyses.

Analysis of the peptides identified with high confidence ($Cn$ > 3.0 and $\Delta Cn$ > 0.2) that were eluted from WT, TAP$^{-/-}$ and ERAAP$^{-/-}$ splenocytes surprisingly revealed little overlap between the peptides displayed by WT cells and either TAP$^{-/-}$ or ERAAP$^{-/-}$ cells (Fig. 2 and Supporting Information Table 1). Only 22.5% of the H2A$^b$-restricted self peptide sequences displayed by WT cells were also presented by TAP$^{-/-}$ or ERAAP$^{-/-}$ cells (Fig. 2A). In a different project, replicate MS samples that consisted of peptides with similar confidence levels eluted from MHC molecules, demonstrated a 63% concordance (SBC, CTS, AJL and SJ, unpublished data). Since class II–associated peptides expressed by WT- and CAP-deficient cells have only 22.5% overlap, the differences in the WT and CAP peptidomes are likely real and not caused by

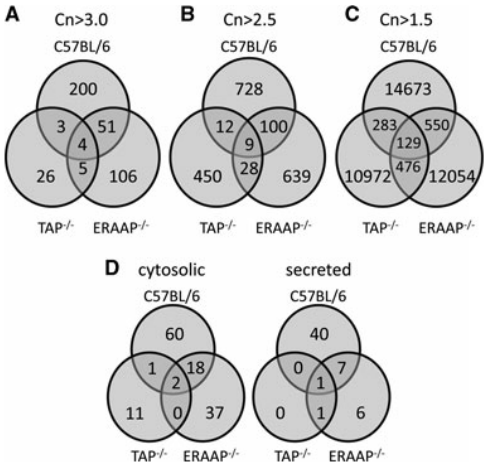



**Figure 2.** TAP and ERAAP deficiency alters the basal H2A$^b$-restricted self peptidome. The prevalence of H2A$^b$-restricted self peptide sequences was compared between WT, TAP$^{-/-}$ and ERAAP$^{-/-}$ strains. Venn diagrams indicate the number of unique and common peptide sequences identified amongst the peptidomes displayed by the indicated strains. $Cn$ > 3.0 (A), $Cn$ > 2.5 (B) or $Cn$ > 1.5 (C) indicates decreasing spectral confidence (see *Materials and Methods*). $\Delta Cn \geq 0.2$ distinguishes between the top two peptide sequences predicted from the spectrum; this criterion allows identification of the best peptide sequence that matches the observed spectrum. (D) Using the LOCATE database, the number of peptides derived from cytosolic and secreted proteins was compared amongst the peptidomes consisting of peptides with $Cn$ > 3.0.

irreproducibility in the experiment. Conversely, 18.4% of self peptide sequences displayed by TAP$^{-/-}$ cells were presented by WT cells, while 33% of self peptide sequences displayed by ERAAP$^{-/-}$ cells were presented by WT cells. This lack of identity was not due to bias in selecting peptides with $Cn$ > 3.0 as datasets which included peptides identified with either moderate ($Cn$ > 2.5 and $\Delta Cn$ > 0.2; Fig. 2B) or low ($Cn$ > 1.5 and $\Delta Cn$ > 0.2; Fig. 2C) confidence also demonstrated little overlap in peptide sequence. However, to maintain focus on relevant naturally processed self peptides using this unbiased approach, all downstream analyses were performed on peptides with $Cn$ > 3.0 and $\Delta Cn$ > 0.2. Importantly, this peptide set was found to have a false discovery rate (FDR; described in the *Materials and Methods*) of 0, i.e. no peptides were identified by random similarity.

Notably, the average length of H2A$^b$-associated peptides increased from 14–16 amino acid residues in WT cells to 18–20 amino acids in TAP$^{-/-}$ and ERAAP$^{-/-}$ cells (Supporting Information Table 1 and Fig. 2). This was consistent with peptide length changes previously observed for class I–associated peptides displayed by ERAAP$^{-/-}$ cells [22]. In addition, we observed numerous groups of nested peptides arising from the same protein (Supporting Information Table 2) as would be expected from class II–associated peptides expressed by WT cells [37, 38]. These nested peptides contained both *N*- and *C*-terminal extensions, consistent with previous reports on class II–associated peptides expressed

by WT cells [37, 38]. Moreover, only two peptides identified in this study have been previously reported (Supporting Information Table 1) [37, 38]. The lack of overlap in peptides identified in previous studies and this one may have resulted from the analysis of different cell populations. We used unmanipulated APCs isolated directly ex vivo in this study compared with B-cell lymphomas, LPS-induced B-cell blasts, IFN-γ–induced BMC2.3 cell line and Flt3-induced cells used in the earlier reports [37, 38]. In addition, although we found thousands of peptides by LC-MS/MS, we have focused solely on those with the highest *Cn* values. It is conceivable that the few hundred peptides previously reported were excluded based on the criteria used for sequence determination and validation and may be present in the larger dataset. Hence the differences observed in the different reports do not detract from the novel peptides reported herein as similar results were observed with the larger datasets as well (Fig. 2B and C).

$H2A^b$-associated peptides were derived from both secreted/extracellular and cytosolic/intracellular proteins as defined in the LOCATE database [39]. However, the majority (~70%) were processed from cytosolic/intracellular proteins (Fig. 2D), including proteins associated with endosomes. Comparing individual genotypes, the presentation of cytoplasmic/intracellular protein-derived peptides was increased in $TAP^{-/-}$ and $ERAAP^{-/-}$ splenocytes. Consistent with previous reports [40], ~63% of the $H2A^b$-associated self peptidome presented by WT cells were generated from cytosolic/intracellular proteins. In contrast, 87.5% and 80.2% of the $H2A^b$-associated peptides displayed by $TAP^{-/-}$ and $ERAAP^{-/-}$ splenocytes, respectively, were derived from cytosolic/intracellular proteins (Fig. 2D). These data demonstrate that numerous cytoplasmic/intracellular proteins, including endosomal proteins, are processed and presented by $H2A^b$ in $TAP^{-/-}$ and $ERAAP^{-/-}$ mice. From these analyses, we conclude that CAP components can impact the $H2A^b$-associated self peptidome.

## TAP and ERAAP deficiency alter the CD4+ TCR repertoire

As the self peptidome instructs the developing TCR repertoire, we compared TCR Vβ usage by $CD4^+$ $CD62L^{HI}CD44^{LO}$ naive T ($T_n$) cells between WT mice and for $TAP^{-/-}$ or $ERAAP^{-/-}$ animals using a panel of Vβ-specific antibodies. As previously reported [35], the frequencies of TCR Vβ usage between WT-, $TAP^{-/-}$- or $ERAAP^{-/-}$-derived $CD4^+$ $T_n$ cells were quite similar, although not identical (Fig. 3A). Likewise, TCR Vβ usage within *Lm*-reactive $CD4^+$ $CD62L^{LO}CD44^{HI}$ effector T ($T_{eff}$) cells of WT, $TAP^{-/-}$ or $ERAAP^{-/-}$ mice were similar as well (Fig. 3B).

Since Ag recognition is mediated by the highly variable CDR3, we specifically examined this region of the TCR β-chains. CDR3β sequence diversity can be estimated by analysing the number of amino acids spanning the V-D-J recombination site by spectratyping the nucleotides that encode them [41, 42]. Although different sequences may have equivalent lengths, thereby underestimating the true diversity, differences in the number of amino acids, nonetheless, provide a high throughput estimate of Ag receptor diversity. The diversity of the TCR of flow sorted $CD4^+$ $T_n$ cells were analysed by spectratyping 52 *Vβ-Jβ* pairings. This analysis revealed extensive alterations in some but not all CDR3β length profiles in the naive TCR β-chain repertoire expressed by WT, $TAP^{-/-}$ or $ERAAP^{-/-}$ mice (Fig. 4 and Supporting Information Fig. 3A). Similar analysis of flow sorted *Lm*-responsive $CD4^+$ $T_{eff}$ cells revealed extensive differences in the CDR3β length profiles between WT and TAP- or ERAAP-deficient $CD4^+$ $T_{eff}$ cells (Fig. 5 and Supporting Information Fig. 3B). These data suggest that, despite similarities in Vβ usage, which was serologically determined, $CD4^+$ T cells utilize different CDR3β sequences in the absence of the CAP machinery. Since the CDR3β region of the TCR is predominantly involved in Ag recognition, sequence differences in this region could potentially lead to alterations in the $CD4^+$ T-cell responses to microbial challenge.

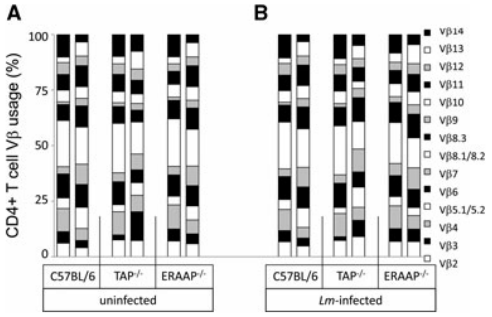



**Figure 3.** Differential self peptidome display has little impact on the TCR Vβ usage. WT, $TAP^{-/-}$ and $ERAAP^{-/-}$ mice were inoculated with *Lm* or not and the TCR Vβ usage of the indicated CD4+ T-cell population was determined by flow cytometry after staining with a panel of Vβ-specific antibodies. The cumulative bar graphs indicate the proportion of each Vβ segment present within the (A) $CD4^+$ $T_n$ ($CD44^{LO}CD62L^{HI}$) or (B) *Lm*-immune $T_{eff}$ ($CD44^{HI}CD62L^{LO}$) population in replicate experiments.

## TAP-deficiency alters class II–restricted microbial Ag recognition

Previously, we reported that the magnitude of the $CD4^+$ T cell response to minor histocompatibility Ag HY and *Lm*-derived LLO and p60 peptides were increased in animals deficient in TAP or ERAAP [21]. Here, we have shown that TAP and ERAAP impact the quality of the $H2A^b$-restricted self peptidome and alter the TCR repertoire. Therefore, we queried whether the CAP machinery could destroy and/or create class II–restricted microbial peptides recognized by $CD4^+$ T cells. To this end, WT, $H2Ab^{-/-}$ and $TAP^{-/-}$ mice were inoculated with VACV and, 7 days later, the TH response tested against a panel of 448 15-mer peptides. This panel consisted of putative $H2A^b$-restricted peptides from VACV ORFs [43]. An initial screen of these peptides revealed few shared specificities and significant alterations in the magnitude of $CD4^+$ T-cell responses to these shared peptides in $TAP^{-/-}$ mice when compared to WT

altered recognition of microbial peptides leading to either limited immunogenicity or enhanced immunopathology. In this regard, it is noteworthy that herpetic stromal keratitis (HSK) — a leading cause of blindness that has an infectious etiology [62] — evolves as a consequence of chronic HSV infection. HSK is a chronic inflammatory disease that is mediated by $CD4^+$ T cells [63]. As ICP47 of HSV blocks TAP function [48], one might predict that the display of an altered peptidome by HSV-infected cells might lead to $CD4^+$ T-cell–mediated inflammation resulting in HSK. Further investigations will be needed to understand the clinical outcome of CAP deficiencies in humans.

In sum, it is becoming clearer that many Th-cell epitopes are being processed by components of both cytosolic and endo/lysosomal Ag-processing pathways [11–15, 21, 27–31, 61]. Data obtained from tagged Ags have suggested that the subcellular localization of the Ag may be critical for its presentation [15, 31, 34, 64–66]. Proteasomes and endo/lysosomal proteases may degrade proteins at the point of Ag entry, endogenous versus exogenous, respectively. Subsequently, peptides may then be shared between the two Ag presentation pathways depending on the efficiency of molecular components that transport processed Ags. While some peptides can be presented by both pathways [11–15, 27–31], it is evident that other peptides are restricted to a single presentation pathway [32, 34]. This is likely due to an as yet undefined biochemical mechanism(s) by which partially processed Ags are targeted from the cytosol to the endo/lysosome. Understanding the underlying mechanism will impact how T-cell biology is harnessed for vaccinations and immunotherapies as well as in treating autoimmune disorders that have a microbial etiology (e.g. HSK).

## Materials and methods

### Animals

C57BL/6J mice were purchased from The Jackson Laboratory. B6.129-$TAP^{-/-}$, B6.129-$ERAAP^{-/-}$ and B6.129-$Ab^{-/-}$ mice [21] were bred, maintained and used in experiments in compliance with Vanderbilt University's Institutional Animal Care and Use Committee regulations and approval. B6.129-$TAP^{-/-}$, B6.129-$ERAAP^{-/-}$ and B6.129-$Ab^{-/-}$ mice had been backcrossed to the C57BL/6 strain 8–10 generations before use.

### Isolation of naturally processed $H2A^b$-associated self peptides

RBC-depleted single cell suspensions of splenocytes pooled from 68 to 70 mice per strain were solubilized, clarified and precleared with normal mouse serum by previously described methods [67, 68]. Pre-cleared lysates were passed twice over protein A Sepharose (Repligen)-bound W6/32 (anti-HLA class I, an irrelevant Ab; Cedarlane) column followed by bead-bound $H2A^b$-specific Ab column (NYRmI-A, Cedarlane) at 4°C. After extensive washes, columns were eluted with 0.2N acetic acid. The eluates were adjusted to 2N acetic acid, incubated for 20 min in a boiling water bath and cooled on ice [68]. Eluted peptides were enriched by Centricon 10 ultrafiltration (Millipore), freeze dried, resuspended in ~0.1 mL deionized distilled water (Sigma) and fractionated by RPC (HP1090, Hewlett-Packard) as described [68]. Approximately 150 fractions were collected and lyophilized to dryness.

### MS-ESI sequencing of naturally processed $H2A^b$-associated self peptides

Each lyophilized RPC fraction was resuspended in 0.1% formic acid and subjected to reversed-phase microcapillary LC-MS/MS analysis using an LTQ linear ion trap mass spectrometer (Thermofisher). A fritless, microcapillary column (100 μm inner diameter) was packed with 10 cm of 5-μm $C_{18}$ reversed-phase material (Synergi 4u Hydro RP80a, Phenomenex). RPC fractionated peptides were loaded onto the column equilibrated in buffer A (0.1% formic acid, 5% acetonitrile) using the LCPacking autosampler. The column was placed in line with an LTQ mass spectrometer. Peptides were eluted using a 60 min linear gradient from 0 to 60% buffer B (0.1% formic acid, 80% acetonitrile) at a flow rate of 0.3 μL/min. During the gradient, the eluted ions were analyzed by one full precursor MS scan (400–2000 $m/z$) followed by five MS/MS scans of the five most abundant ions detected in the precursor MS scan while operating under dynamic exclusion. Extractms2 program was used to generate the ASCII peak list and to identify +1 or multiply charged precursor ions from the native mass spectrometry data file [69]. Tandem spectra were searched with no protease specificity using SEQUEST-PVM against a RefSeq murine protein database [36]. For multiply charged precursor ions ($z \geq +2$), an independent search was performed on both the +2 and +3 mass of the parent ion. Data were processed and organized using the BIGCAT software analysis suite with a weighted scoring matrix used to select the most likely charge state of multiply charged precursor ions [70]. Fragmentation/ionization patterns were compared against a dataset consisting of the fragmentation/ionization patterns of theoretical and known peptide sequences. The degree of concordance between these two patterns was assigned a cross correlation score $X_{corr}$ ($Cn$) with higher values representing greater concordance between the observed and expected fragmentation/ionization patterns [36]. Peptides with a Sequest $Cn$ score >3.0 and $\Delta Cn > 0.2$ compared with the second most likely assignment were considered highly concordant (see Supporting Information Fig. 1).

The ion fragments were also searched against the reversed mouse proteome database to determine the false detection rate FDR. FDR was calculated as (2 × # reverse hits)/(# reverse hits + # forward hits). This generated an overall FDR of 7%. Whereas a search of only the highly concordant peptide spectra ($Cn > 3.0$ and $\Delta Cn > 0.2$) generated an FDR of 0, i.e. no peptides were identified in the reversed database. The parental ions representing peptides eluted from class II molecules of only two genotypes

were manually searched against the database of parental ions of the third genotype. Of the 62 overlapping peptide sequences, only 2 (3.2%) were identified in the third genotype within 10 HPLC fractions and 10 min of LC elution of the same fraction number/retention time. Of these, one was inappropriately identified by the tandem MS and the other was not analysed by tandem MS for identification. From this analysis, we conclude that 96.8% of peptides presented by class II molecules of only two genotypes were correctly identified and were not presented by that of the third genotype.

### Immunization, T-cell purification and functional analysis

The indicated mouse strains were inoculated either retro-orbitally with $\sim 5 \times 10^4$ cfu WT *Lm* or i.p. with $2 \times 10^5$ pfu VACV WR strain. After 7 days, splenocytes were harvested and either stained for flow cytometric characterization or restimulated for functional analyses. *Lm*-immune splenocytes were stained with mAb against mouse CD62L and CD44 for flow sorting of naive ($T_n$) and effector ($T_{eff}$) $CD4^+$ T-cell populations (FACS Aria, BD Bioscience). Post-sort purity was ascertained by flow cytometry and found to be >98% (data not shown). A separate aliquot of $CD4^+$ T cells were analysed for Vβ usage with a panel of 15 anti-Vβ antibodies (BD Bioscience) within the $T_n$ ($CD44^{lo}CD62L^{hi}$) or *Lm*-immune $T_{eff}$ ($CD44^{hi}CD62L^{lo}$) subsets.

Co-culture of total VACV-immune splenocytes with $H2A^b$-restricted peptides derived from VACV [43] for IFN-γ ELISPOT was performed as previously described [21].

### TCR spectratyping

Total RNA was isolated from flow sorted non-immune $CD4^+$ T cells or flow sorted $CD62L^{hi}CD44^{lo}CD4^+$ $T_n$ cells and activated, effector $CD62L^{lo}CD44^{hi}CD4^+$ $T_{eff}$ cells and converted to cDNA as described [71]. PCR amplification of individual *V*β-*C*β junctions and *J*β-specific run-off was performed using previously reported primer pairs [72] and Supermix (Invitrogen). The run-off *J*β primers were end-modified with WellRED D2, D3 or D4 fluorescent dyes (Sigma-Genosys) to detect products using capillary gel electrophoresis (CEQ8000; Beckman Coutler). CDR3β fragment sizes were determined by correlation against a size standard consisting of WellRED D1 fluorescent DNA strands of incremental 20 nt residues (Beckman-Coulter) and the frequency within the population was determined by integration of the peak area. CDR3β length was calculated as the number of amino acids between the conserved last germline encoded Vβ Cys to the Jβ Gly-X-Gly motif.

Acknowledgements: This work was supported by grant G12MD007592 from the National Institutes on Minority Health and Health Disparities (NIMHD), a component of the National Institutes of Health (NIH) as well as training (HL069765), research (HL054977 and AI040079 to S.J. and AI040024 to A.S.) and core (CA068485 & DK058404) grants from the NIH.

Conflicts of interest: The authors declare no financial or commercial conflict of interest.

## References

1 **Cresswell, P., Ackerman, A. L., Giodini, A., Peaper, D. R. and Wearsch, P. A.**, Mechanisms of MHC class I-restricted antigen processing and cross-presentation. *Immunol. Rev.* 2005. **207**: 145–157.

2 **Bryant, P. and Ploegh, H.**, Class II MHC peptide loading by the professionals. *Curr. Opin. Immunol.* 2004. **16**: 96–102.

3 **Lee, H. K., Mattei, L. M., Steinberg, B. E., Alberts, P., Lee, Y. H., Chervonsky, A., Mizushima, N. et al.**, In vivo requirement for Atg5 in antigen presentation by dendritic cells. *Immunity* 2011. **32**: 227–239.

4 **Nimmerjahn, F., Milosevic, S., Behrends, U., Jaffee, E. M., Pardoll, D. M., Bornkamm, G. W. and Mautner, J.**, Major histocompatibility complex class II-restricted presentation of a cytosolic antigen by autophagy. *Eur. J. Immunol.* 2003. **33**: 1250–1259.

5 **Paludan, C., Schmid, D., Landthaler, M., Vockerodt, M., Kube, D., Tuschl, T. and Munz, C.**, Endogenous MHC class II processing of a viral nuclear antigen after autophagy. *Science* 2005. **307**: 593–596.

6 **Munz, C.**, Selective macroautophagy for immunity. *Immunity* 2010. **32**: 298–299.

7 **Schmid, D., Pypaert, M. and Munz, C.**, Antigen-loading compartments for major histocompatibility complex class II molecules continuously receive input from autophagosomes. *Immunity* 2007. **26**: 79–92.

8 **Zhou, D., Li, P., Lin, Y., Lott, J. M., Hislop, A. D., Canaday, D. H., Brutkiewicz, R. R. et al.**, Lamp-2a facilitates MHC class II presentation of cytoplasmic antigens. *Immunity* 2005. **22**: 571–581.

9 **Crotzer, V. L. and Blum, J. S.**, Autophagy and its role in MHC-mediated antigen presentation. *J. Immunol.* 2009. **182**: 3335–3341.

10 **Crotzer, V. L. and Blum, J. S.**, Cytosol to lysosome transport of intracellular antigens during immune surveillance. *Traffic* 2008. **9**: 10–16.

11 **Bonifaz, L. C., Arzate, S. and Moreno, J.**, Endogenous and exogenous forms of the same antigen are processed from different pools to bind MHC class II molecules in endocytic compartments. *Eur. J. Immunol.* 1999. **29**: 119–131.

12 **Jacobson, S., Sekaly, R. P., Jacobson, C. L., McFarland, H. F. and Long, E. O.**, HLA class II-restricted presentation of cytoplasmic measles virus antigens to cytotoxic T cells. *J. Virol.* 1989. **63**: 1756–1762.

13 **Lich, J. D., Elliott, J. F. and Blum, J. S.**, Cytoplasmic processing is a prerequisite for presentation of an endogenous antigen by major histocompatibility complex class II proteins. *J. Exp. Med.* 2000. **191**: 1513–1524.

14 **Mukherjee, P., Dani, A., Bhatia, S., Singh, N., Rudensky, A. Y., George, A., Bal, V. et al.**, Efficient presentation of both cytosolic and endogenous transmembrane protein antigens on MHC class II is dependent on cytoplasmic proteolysis. *J. Immunol.* 2001. **167**: 2632–2641.

15 **Dani, A., Chaudhry, A., Mukherjee, P., Rajagopal, D., Bhatia, S., George, A., Bal, V. et al.**, The pathway for MHCII-mediated presentation of endogenous proteins involves peptide transport to the endo-lysosomal compartment. *J. Cell Sci.* 2004. **117**: 4219–4230.

16 **Jaraquemada, D., Marti, M. and Long, E. O.**, An endogenous processing pathway in vaccinia virus-infected cells for presentation of cytoplasmic antigens to class II-restricted T cells. *J. Exp. Med.* 1990. **172**: 947–954.

17 **Malnati, M. S., Marti, M., LaVaute, T., Jaraquemada, D., Biddison, W., DeMars, R. and Long, E. O.**, Processing pathways for presentation of cytosolic antigen to MHC class II-restricted T cells. *Nature* 1992. **357**: 702–704.

18 **Reed, A. J., Noorchashm, H., Rostami, S. Y., Zarrabi, Y., Perate, A. R., Jeganathan, A. N., Caton, A. J. et al.**, Alloreactive CD4 T-cell activation in vivo: an autonomous function of the indirect pathway of alloantigen presentation. *J. Immunol.* 2003. **171**: 6502–6509.

19 **Richards, D. M., Dalheimer, S. L., Ehst, B. D., Vanasek, T. L., Jenkins, M. K., Hertz, M. I. and Mueller, D. L.**, Indirect minor histocompatibility antigen presentation by allograft recipient cells in the draining lymph node leads to the activation and clonal expansion of CD4+ T cells that cause obliterative airways disease. *J. Immunol.* 2004. **172**: 3469–3479.

20 **Chen, Y., Demir, Y., Valujskikh, A. and Heeger, P. S.**, The male minor transplantation antigen preferentially activates recipient CD4+ T cells through the indirect presentation pathway in vivo. *J. Immunol.* 2003. **171**: 6510–6518.

21 **Dragovic, S. M., Hill, T., Christianson, G. J., Kim, S., Elliott, T., Scott, D., Roopenian, D. C. et al.**, Proteasomes, TAP, and endoplasmic reticulum-associated aminopeptidase associated with antigen processing control CD4+ Th cell responses by regulating indirect presentation of MHC class II-restricted cytoplasmic antigens. *J. Immunol.* 2011. **186**: 6683–6692.

22 **Blanchard, N., Kanaseki, T., Escobar, H., Delebecque, F., Nagarajan, N. A., Reyes-Vargas, E., Crockett, D. K. et al.**, Endoplasmic reticulum aminopeptidase associated with antigen processing defines the composition and structure of MHC class I peptide repertoire in normal and virus-infected cells. *J. Immunol.* 2010. **184**: 3033–3042.

23 **Hammer, G. E., Gonzalez, F., James, E., Nolla, H. and Shastri, N.**, In the absence of aminopeptidase ERAAP, MHC class I molecules present many unstable and highly immunogenic peptides. *Nat. Immunol.* 2007. **8**: 101–108.

24 **Hammer, G. E., Gonzalez, F., Champsaur, M., Cado, D. and Shastri, N.**, The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules. *Nat. Immunol.* 2006. **7**: 103–112.

25 **Van, K. L., shton-Rickardt, P. G., Ploegh, H. L. and Tonegawa, S.**, TAP1 mutant mice are deficient in antigen presentation, surface class I molecules, and CD4–8+ T cells. *Cell* 1992. **71**: 1205–1214.

26 **Yan, J., Parekh, V. V., Mendez-Fernandez, Y., Olivares-Villagomez, D., Dragovic, S., Hill, T., Roopenian, D. C. et al.**, In vivo role of ER-associated peptidase activity in tailoring peptides for presentation by MHC class Ia and class Ib molecules. *J. Exp. Med.* 2006. **203**: 647–659.

27 **Tewari, M. K., Sinnathamby, G., Rajagopal, D. and Eisenlohr, L. C.**, A cytosolic pathway for MHC class II-restricted antigen processing that is proteasome and TAP dependent. *Nat. Immunol.* 2005. **6**: 287–294.

28 **Carmichael, P., Kerr, L.-A., Kelly, A., Lombardi, G., Zeigler, B. U., Ziegler, A., Trowsdale, J. et al.**, The TAP complex influences allorecognition of class II MHC molecules. *Hum. Immunol.* 1996. **50**: 70–77.

29 **Gueguen, M. and Long, E. O.**, Presentation of a cytosolic antigen by major histocompatibility complex class II molecules requires a long-lived form of the antigen. *Proc. Natl. Acad. Sci. U.S.A.* 1996. **93**: 14692–14697.

30 **Goldszmid, R. S., Bafica, A., Jankovic, D., Feng, C. G., Caspar, P., Winkler-Pickett, R., Trinchieri, G. et al.**, TAP-1 indirectly regulates CD4+ T-cell priming in Toxoplasma gondii infection by controlling NK cell IFN-gamma production. *J. Exp. Med.* 2007. **204**: 2591–2602.

31 **Malnati, M. S., Ceman, S., Weston, M., DeMars, R. and Long, E. O.**, Presentation of cytosolic antigen by HLA-DR requires a function encoded in the class II region of the MHC. *J. Immunol.* 1993. **151**: 6751–6756.

32 **Dissanayake, S. K., Tuera, N. and Ostrand-Rosenberg, S.**, Presentation of endogenously synthesized MHC class II-restricted epitopes by MHC class II cancer vaccines is independent of transporter associated with Ag processing and the proteasome. *J. Immunol.* 2005. **174**: 1811–1819.

33 **Loss, G. E., Jr., Elias, C. G., Fields, P. E., Ribaudo, R. K., McKisic, M. and Sant, A. J.**, Major histocompatibility complex class II-restricted presentation of an internally synthesized antigen displays cell-type variability and segregates from the exogenous class II and endogenous class I presentation pathways. *J. Exp. Med.* 1993. **178**: 73–85.

34 **Oxenius, A., Bachmann, M. F., Ashton-Rickardt, P. G., Tonegawa, S., Zinkernagel, R. M. and Hengartner, H.**, Presentation of endogenous viral proteins in association with major histocompatibility complex class II: on the role of intracellular compartmentalization, invariant chain and the TAP transporter system. *Eur. J. Immunol.* 1995. **25**: 3402–3411.

35 **Tourne, S., van Santen, H. M., van Roon, M., Berns, A., Benoist, C., Mathis, D. and Ploegh, H.**, Biosynthesis of major histocompatibility complex molecules and generation of T cells in Ii TAP1 double-mutant mice. *Proc. Natl. Acad. Sci. U.S.A.* 1996. **93**: 1464–1469.

36 **Eng, J. K., McCormack, A. L. and Yates, J. R.**, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994. **5**: 976–989.

37 **Bozzacco, L., Yu, H., Zebroski, H. A., Dengjel, J., Deng, H., Mojsov, S. and Steinman, R. M.**, Mass spectrometry analysis and quantitation of peptides presented on the MHC II molecules of mouse spleen dendritic cells. *J. Proteome Res.* 2011. **10**: 5016–5030.

38 **Dongre, A. R., Kovats, S., deRoos, P., McCormack, A. L., Nakagawa, T., Paharkova-Vatchkova, V., Eng, J. et al.**, In vivo MHC class II presentation of cytosolic proteins revealed by rapid automated tandem mass spectrometry and functional analyses. *Eur. J. Immunol.* 2001. **31**: 1485–1494.

39 **Fink, J. L., Aturaliya, R. N., Davis, M. J., Zhang, F., Hanson, K., Teasdale, M. S., Kai, C. et al.**, LOCATE: a mouse protein subcellular localization database. *Nucleic Acids Res.* **34**: D213–D217.

40 **Rudensky, A., Preston-Hurlburt, P., Hong, S. C., Barlow, A. and Janeway, C. A., Jr.**, Sequence analysis of peptides bound to MHC class II molecules. *Nature* 1991. **353**: 622–627.

41 **Spencer, C. T., Abate, G., Blazevic, A. and Hoft, D. F.**, Only a subset of phosphoantigen-responsive gamma9delta2 T cells mediate protective tuberculosis immunity. *J. Immunol.* 2008. **181**: 4471–4484.

42 **Gregersen, P. K., Hingorani, R. and Monteiro, J.**, Oligoclonality in the CD8+ T-cell population. *Ann. N.Y. Acad. Sci.* 1995. **756**: 19–27.

43 **Moutaftsi, M., Bui, H. H., Peters, B., Sidney, J., Salek-Ardakani, S., Oseroff, C., Pasquetto, V. et al.**, Vaccinia virus-specific CD4 +T-cell responses target a set of antigens largely distinct from those targeted by CD8+ T-cell responses. *J. Immunol.* 2007. **178**: 6814–6820.

44 **Halenius, A., Hauka, S., Dolken, L., Stindt, J., Reinhard, H., Wiek, C., Hanenberg, H. et al.**, Human cytomegalovirus disrupts the major histocompatibility complex class I peptide-loading complex and inhibits tapasin gene transcription. *J. Virol.* 2011. **85**: 3473–3485.

45 **Horst, D., Favaloro, V., Vilardi, F., van Leeuwen, H. C., Garstka, M. A., Hislop, A. D., Rabu, C. et al.**, EBV protein BNLF2a exploits host tail-anchored protein integration machinery to inhibit TAP. *J. Immunol.* 2011. **186**: 3594–3605.

46 **Kasajima, A., Sers, C., Sasano, H., Johrens, K., Stenzinger, A., Noske, A., Buckendahl, A. C. et al.**, Down-regulation of the antigen processing machinery is linked to a loss of inflammatory response in colorectal cancer. *Hum. Pathol.* 2010. **41**: 1758–1769.

47 Verweij, M. C., Lipinska, A. D., Koppers-Lalic, D., van Leeuwen, W. F., Cohen, J. I., Kinchington, P. R., Messaoudi, I. et al., The capacity of UL49.5 proteins to inhibit TAP is widely distributed among members of the genus Varicellovirus. *J. Virol.* 2011. **85**: 2351–2363.

48 Verweij, M. C., Ressing, M. E., Knetsch, W., Quinten, E., Halenius, A., van Bel, N., Hengel, H. et al., Inhibition of mouse TAP by immune evasion molecules encoded by non-murine herpesviruses. *Mol. Immunol.* 2011. **48**: 835–845.

49 Einstein, M. H., Leanza, S., Chiu, L. G., Schlecht, N. F., Goldberg, G. L., Steinberg, B. M. and Burk, R. D., Genetic variants in TAP are associated with high-grade cervical neoplasia. *Clin. Cancer Res.* 2009. **15**: 1019–1023.

50 de la Salle, H., Hanau, D., Fricker, D., Urlacher, A., Kelly, A., Salamero, J., Powis, S. H. et al., Homozygous human TAP peptide transporter mutation in HLA class I deficiency. *Science* 1994. **265**: 237–241.

51 Matamoros, N., Mila, J., Llano, M., Balas, A., Vicario, J. L., Pons, J., Crespi, C. et al., Molecular studies and NK cell function of a new case of TAP2 homozygous human deficiency. *Clin. Exp. Immunol.* 2001. **125**: 274–282.

52 Colonna, M., Bresnahan, M., Bahram, S., Strominger, J. L. and Spies, T., Allelic variants of the human putative peptide transporter involved in antigen processing. *Proc. Natl. Acad. Sci. U.S.A.* 1992. **89**: 3932–3936.

53 Mach, B., Steimle, V. and Reith, W., MHC class II-deficient combined immunodeficiency: a disease of gene regulation. *Immunol. Rev.* 1994. **138**: 207–221.

54 Will, N., Seger, R. A., Betzler, C., Dockter, G., Graf, N., Büttner, M., Irlé, C. et al., Bare lymphocyte syndrome: combined immunodeficiency and neutrophil dysfunction. *Eur. J. Pediatr.* 1990. **149**: 700–704.

55 Douhan, J., Hauber, I., Eibl, M. M. and Glimcher, L. H., Genetic evidence for a new type of major histocompatibility complex class II combined immunodeficiency characterized by a dyscoordinate regulation of HLA-D alpha and beta chains. *J. Exp. Med.* 1996. **183**: 1063–1069.

56 Kovats, S., Nepom, G. T., Coleman, M., Nepom, B., Kwok, W. W. and Blum, J. S., Deficient antigen-presenting cell function in multiple genetic complementation groups of type II bare lymphocyte syndrome. *J. Clin. Invest.* 1995. **96**: 217–223.

57 DeSandro, A., Nagarajan, U. M. and Boss, J. M., The bare lymphocyte syndrome: molecular clues to the transcriptional regulation of major histocompatibility complex class II genes. *Am. J. Hum. Genet.* 1999. **65**: 279–286.

58 Mehta, A. M., Jordanova, E. S., Corver, W. E., van Wezel, T., Uh, H.-W., Kenter, G. G. and Jan Fleuren, G., Single nucleotide polymorphisms in antigen processing machinery component ERAP1 significantly associate with clinical outcome in cervical carcinoma. *Gene Chromosome Canc.* 2009. **48**: 410–418.

59 Evnouchidou, I., Birtley, J., Seregin, S., Papakyriakou, A., Zervoudi, E., Samiotaki, M., Panayotou, G. et al., A common single nucleotide polymorphism in endoplasmic reticulum aminopeptidase 2 induces a specificity switch that leads to altered antigen processing. *J. Immunol.* 2012. **189**: 2383–2392.

60 Johnson, M., Roten, L., Dyer, T., East, C., Forsmo, S., Blangero, J., Brennecke, S. et al., The ERAP2 gene is associated with preeclampsia in Australian and Norwegian populations. *Hum. Genet.* 2009. **126**: 655–666.

61 Guillaume, B., Stroobant, V., Bousquet-Dubouch, M. P., Colau, D., Chapiro, J., Parmentier, N., Dalet, A. et al., Analysis of the processing of seven human tumor antigens by intermediate proteasomes. *J. Immunol.* 2012. **189**: 3538–3547.

62 Deshpande, S., Banerjee, K., Biswas, P. S. and Rouse, B. T., Herpetic eye disease: immunopathogenesis and therapeutic measures. *Expert Rev. Mol. Med.* 2004. **6**: 1–14.

63 Banerjee, K., Biswas, P. S. and Rouse, B. T., Elucidating the protective and pathologic T-cell species in the virus-induced corneal immunoinflammatory condition herpetic stromal keratitis. *J. Leukoc. Biol.* 2005. **77**: 24–32.

64 Demirel, O., Waibler, Z., Kalinke, U., Grunebach, F., Appel, S., Brossart, P., Hasilik, A. et al., Identification of a lysosomal peptide transport system induced during dendritic cell development. *J. Biol. Chem.* 2007. **282**: 37836–37843.

65 Busch, R., Vturina, I. Y., Drexler, J., Momburg, F. and Hammerling, G. J., Poor loading of major histocompatibility complex class II molecules with endogenously synthesized short peptides in the absence of invariant chain. *Eur. J. Immunol.* 1995. **25**: 48–53.

66 Rudensky, A. Y., Maric, M., Eastman, S., Shoemaker, L., DeRoos, P. C. and Blum, J. S., Intracellular assembly and transport of endogenous peptide-MHC class II complexes. *Immunity* 1994. **1**: 585–594.

67 Marrack, P., Ignatowicz, L., Kappler, J. W., Boymel, J. and Freed, J. H., Comparison of peptides bound to spleen and thymus class II. *J. Exp. Med.* 1993. **178**: 2173–2183.

68 Joyce, S., Kuzushima, K., Kepecs, G., Angeletti, R. H. and Nathenson, S. G., Characterization of an incompletely assembled major histocompatibility class I molecule (H-2Kb) associated with unusually long peptides: implications for antigen processing and presentation. *Proc. Natl. Acad. Sci. U.S.A.* 1994. **91**: 4145–4149.

69 Gerbasi, V. R. and Link, A. J., The myotonic dystrophy type 2 protein ZNF9 is part of an ITAF complex that promotes cap-independent translation. *Mol. Cell. Proteomics* 2007. **6**: 1049–1058.

70 McAfee, K. J., Duncan, D. T., Assink, M. and Link, A. J., Analyzing proteomes and protein function using graphical comparative analysis of tandem mass spectrometry results. *Mol. Cell. Proteomics* 2006. **5**: 1497–1513.

71 Gordy, L. E., Bezbradica, J. S., Flyak, A. I., Spencer, C. T., Dunkle, A., Sun, J., Stanic, A. K. et al., IL-15 regulates homeostasis and terminal maturation of NKT cells. *J. Immunol.* 2011. **187**: 6335–6345.

72 Pannetier, C., Cochet, M., Darche, S., Casrouge, A., Zöller, M. and Kourilsky, P., The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. *Proc. Natl. Acad. Sci. U.S.A.* 1993. **90**: 4319–4323.

*Abbreviations:* **CAP**: MHC class I antigen processing · **FDR**: false detection rate · **LM**: *Listeria monocytogenes* · **RPC**: reversed-phase chromatography · **VACV**: vaccinia viral

*Full correspondence:* Dr. Charles T. Spencer, 500 W. University Avenue, El Paso, TX 79968, USA
Fax: +1-915-747-5808
e-mail: ctspencer@utep.edu

*Additional correspondence:* Sebastian Joyce, 1161 21st Avenue South, Nashville, TN 37232, USA
Fax: +1-615-343-7392
e-mail: sebastian.joyce@vanderbilt.edu

*Current address:* Charles T. Spencer, Department of Biological Sciences, The University of Texas at El Paso, 500 W. University Avenue, El Paso, TX, USA

Received: 23/10/2012
Revised: 2/1/2013
Accepted: 30/1/2013
Accepted article online: 5/2/2013

# Appendix Z – Manuscript – 5: A Novel Algorithm for Validating Peptide Identification from a Shotgun Proteomics Search Engine

Journal of proteome research

Article

pubs.acs.org/jpr

# A Novel Algorithm for Validating Peptide Identification from a Shotgun Proteomics Search Engine

Ling Jian,[||,#] Xinnan Niu,[#] Zhonghang Xia,[⊥] Parimal Samir,[¶] Chiranthani Sumanasekera,[‡] Zheng Mu, Jennifer L. Jennings, Kristen L. Hoek, Tara Allos, Leigh M. Howard,[§] Kathryn M. Edwards,[§] P. Anthony Weil,[‡] and Andrew J. Link*[,†]

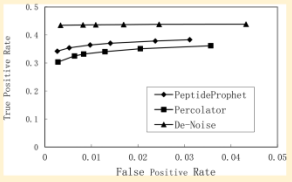
[†]Department of Pathology, Microbiology and Immunology, [‡]Department of Molecular Physiology and Biophysics, [§]Department of Medicine, and [¶]Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, United States

[||]School of Mathematical Sciences, Dalian University of Technology, Dalian, China

[⊥]Department of Mathematics and Computer Science, Western Kentucky University, Bowling Green, Kentucky 42101, United States

Supporting Information

**ABSTRACT:** Liquid chromatography coupled with tandem mass spectrometry (LC−MS/MS) has revolutionized the proteomics analysis of complexes, cells, and tissues. In a typical proteomic analysis, the tandem mass spectra from a LC−MS/MS experiment are assigned to a peptide by a search engine that compares the experimental MS/MS peptide data to theoretical peptide sequences in a protein database. The peptide spectra matches are then used to infer a list of identified proteins in the original sample. However, the search engines often fail to distinguish between correct and incorrect peptides assignments. In this study, we designed and implemented a novel algorithm called De-Noise to reduce the number of incorrect peptide matches and maximize the number of correct peptides at a fixed false discovery rate using a minimal number of scoring outputs from the SEQUEST search engine. The novel algorithm uses a three-step process: data cleaning, data refining through a SVM-based decision function, and a final data refining step based on proteolytic peptide patterns. Using proteomics data generated on different types of mass spectrometers, we optimized the De-Noise algorithm on the basis of the resolution and mass accuracy of the mass spectrometer employed in the LC−MS/MS experiment. Our results demonstrate De-Noise improves peptide identification compared to other methods used to process the peptide sequence matches assigned by SEQUEST. Because De-Noise uses a limited number of scoring attributes, it can be easily implemented with other search engines.




**KEYWORDS:** *proteomics, mass spectrometry, bioinformatics, support vector machines, peptide spectrum match, database search engine, validation*

## INTRODUCTION

Liquid chromatography coupled with tandem mass spectrometry (LC−MS/MS) offers the promise to comprehensively identify and quantify the proteome of complexes, cells, and tissues. The large numbers of peptide spectra generated from LC−MS/MS experiments are routinely searched using a search engine against theoretical fragmentation spectra derived from target databases containing either protein or translated nucleic acid sequences. It is typically assumed that a peptide spectrum match (PSM) for each MS/MS spectrum is contained in the sequence database. In a typical peptide identification procedure, PSMs are ranked according to either a cross correlation, a statistical score, or a probability that the match between the experimental and theoretical is correct and unique. Only those PSMs with the highest scores or most significant probabilities are reported as correct. However, this approach often falsely identifies the peptides. In reality, more than 50% of PSMs initially assigned by database search engines, such as SEQUEST, MASCOT, and X!TANDEM, are incorrect.[1,2] As a result, the accuracy of database search results is often evaluated by searching a decoy protein database to identify the false discovery rate (FDR).[2−8] Decoy databases contain either reversed or randomly shuffled protein sequences derived from the target protein database. The database search engine assigns an observed spectrum to either a target or a decoy sequence. The assignment of a peptide from a decoy database to an experimental spectrum is considered incorrect because it is assumed that there is no such peptide sequence in reality. The target-decoy database search also indicates the quality or reliability of the target PSMs. Nonetheless, the target PSMs are not all correct as a result of either the poor quality of the experimental MS/MS data, the absence of the sequence in the database, or unexpected amino acid modifications. As a consequence, a fraction of the target PSMs from the search engine is false positive. Hence, manual or computational approaches are essential to validate target PSMs after a search engine-protein database analysis of LC−MS/MS data.

SEQUEST is one of the most widely used approaches for automatically assigning the observed spectra generated from a

**Received:** July 12, 2012
**Published:** February 12, 2013





dx.doi.org/10.1021/pr300631t | J. Proteome Res. 2013, 12, 1108−1119

LC−MS/MS experiment to peptide sequences in a sequence database.[9] However, the original SEQUEST algorithm does not include a statistical method to determine the specificity of peptide-spectrum matching. An early approach to identify correct target PSMs uses empirical score filters set at defined score thresholds to validate PSMs from a SEQUEST search.[10,11] PSMs above the defined thresholds are accepted as correct, while those below are assumed to be incorrect. The empirical score filters are not always easily defined due to the multiple scoring metrics derived from SEQUEST scores and the variable quality of the mass spectrometry data. Also, the accuracy of the validated PSMs derived from an empirical scoring filter varies with the type of mass spectrometer used.

Different approaches have been developed to validate peptide assignments.[12] One of the most commonly used computational tools is PeptideProphet, which uses a Bayesian statistical algorithm to convert SEQUEST scores into probabilities.[13] With PeptideProphet, conditional probabilities for the PSMs are computed by the expectation maximization (EM) method, using the assumption that the PSM data are drawn from a mixture in which the distribution of the correct and incorrect PSMs follows a prescribed Gaussian model. A list of PSMs above a predefined posterior probability is reported.[13] An updated version of PeptideProphet utilizing a semi-supervised technique was recently developed.[14] It integrates the EM algorithm with a decoy database search strategy to build a classifier based on a Bayesian probability model.

To provide a more efficient way of evaluating SEQUEST outputs, MacCoss and Noble's group employed support vector machines (SVM), a powerful classification technique,[16,17] to classify correct and incorrect PSMs after a database search.[15−17] SVM-based classification is a supervised learning approach that uses training data to build a model and assign a label to each data point. They used this approach to create the algorithm called Percolator to directly distinguish correct from incorrect PSMs.[4] The goal of Percolator is to increase the number of correct target PSMs reported at a minimal FDR or q-value.[18] Starting with a small set of trusted correct PSMs and a set of incorrect PSMs from searching a decoy database, Percolator iteratively adjusts the learning model to fit the data set by ranking high-confidence PSMs higher than decoy peptide matches. With the given $q$-value, this approach iteratively trains the classifier and eventually results in a classifier that has an improved ability to distinguish correct from incorrect PSMs.

In this study, we have developed a novel algorithm called De-Noise for statistical validation of correct target PSMs identified by SEQUEST. De-Noise uses a SEQUEST search of a concatenated database containing both target and decoy proteins. It uses the decoy PSMs as incorrect references for measuring the reliability of the correct target PSMs. The De-Noise algorithm is a continuous refining process by which the incorrect target PSMs or noisy PSMs are sequentially eliminated. First, it computes the distance of every target PSM to the centroid of the decoy PSMs. With the assumption that the target PSMs close to the centroid of the decoy PSMs are incorrect or noise, they are eliminated on the basis of a defined ratio. The remaining data set provides a set of PSMs with improved quality for building an SVM-based decision function to refine the target and decoy PSM. Using a given false positive rate (FPR), De-Noise distinguishes the correct from incorrect target PSMs using two rounds of SVM-based decision functions and refinement. Specifically, the lowest scoring target PSMs are discarded from the data set on the basis of the scores derived from SVM-based decision functions until the FPR is reached. Next, the algorithm sorts the remaining PSMs based on the expected protease digestion patterns into three categories: canonical, half-canonical, and non-canonical. It assigns the protease digestion categories an expectation factor based upon the expected distribution of the three categories. With the expectation factor and a score ($PSM_{evaluator}$) derived from normalized SEQUEST's *Xcorr* and *DeltaCn* scores, De-Noise further refines the PSMs to eliminate the incorrect target PSMs.

Our results demonstrate De-Noise has increased sensitivity and specificity for validating PSMs after a SEQUEST search compared to both PeptideProphet and Percolator. To evaluate the performance of De-Noise, we used LC−MS/MS data sets generated from various control and biological samples run on different mass spectrometers. The mass spectrometers had a wide range of mass accuracies, resolutions, and user-defined capabilities for selecting the precursor ions to fragment. The low and high data quality data sets were used to develop and evaluate our De-Noise algorithm. Our results demonstrate that De-Noise validates more correct target PSMs under a series of fixed FDRs compared to PeptideProphet and Percolator. The target PSMs validated by these two algorithms extensively overlap De-Noise's validated PSMs. These results demonstrate De-Noise can increase the number of validated target PSMs.

## MATERIALS AND METHODS

### Reagents

Universal Proteomics Standard Set (UPS1) was purchased from Sigma (St. Louis, MO). Partisphere strong cation exchange (SCX) material was purchased from Whatman International Ltd. Jupiter 3 μm C18 300A, reverse phase (RP) material was purchased from Phenomenex (Torrance, CA). Formic acid and HPLC-grade acetonitrile were obtained from Fisher Scientific (Pittsburgh, PA). Trypsin was purchased from Promega (Madison, WI). PEEK tubing, sleeves, microtee, and microcross were obtained from Upchurch Scientific (Oak Harbor, WA). Fused-silica capillaries were purchased from Polymicro Technologies (Phoenix, AZ).

### Sample Preparation and LC−MS/MS Analysis

**UPS1.** UPS1 was solubilized in water, reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAA), and trypsin-digested as previously described.[19] The tryptic peptides were analyzed with RP microcapillary LC−nanoESI/MS/MS. Briefly, a fritless, microcapillary 100-μm i.d. column was packed with 9 cm of Jupiter C18 RP material. A 0.5 pmol aliquot of the trypsin-digested UPS1 was loaded onto the RP column equilibrated in buffer A (0.1% formic acid, 5% acetonitrile). The column was placed in line with an LTQ linear ion trap mass spectrometer (ThermoFisher). The sample was eluted using a 60-min linear gradient from 0 to 60% buffer B (0.1% formic acid, 80% acetonitrile) at a flow rate of 0.3 μL/min. During the gradient, the eluted ions were analyzed by one full precursor MS scan (400−2000 $m/z$) followed by MS/MS scans on the five most intense precursor ions detected in the precursor MS scan while operating under dynamic exclusion.

**Gcn4.** Affinity preparation of the *S. cerevisiae* Gcn4 complex and its MudPIT analysis using an LCQ quadrupole ion trap mass spectrometer (Thermofisher) have been previously described.[19]

**Tal08.** *S. cerevisiae* transcription complexes were prepared from yeast YTT3675 cells using the Tal08 minichromosome[20] and Dynal beads (Invitrogen) cross-linked to anti-Flag M2 antibody (Sumanasekera et al., manuscript in preparation). For Tal08

complexes, 20 ng of BSA was added to the Tal08 sample as an internal standard. The Tal08 sample was reduced with DTT, alkylated with IAA, and trypsin-digested as previously described.[19] The desalted tryptic peptides were analyzed using a 11-step MudPIT experiment on a LTQ-Orbitrap XL (ThermoFisher). A 100-$\mu$M i.d. fused silica microcapillary packed with 3 cm of PartiSphere SCX material was coupled to a 100-$\mu$M fritless fused silica microcapillary column packed with 12 cm Jupiter C18 RP material. For the Mudpit run, the salt steps were 25 mM, 50 mM, 75 mM, 100 mM, 150 mM, 200 mM, 250 mM, 300 mM, 500 mM, 750 mM, and 1 M ammonium acetate. A 100 min linear RP gradient from 5% to 45% Buffer B was used for each salt step. Buffer A was 0.1% formic acid in HPLC-grade water, and Buffer B was 0.1% formic acid in acetonitrile. A precursor ion scan was performed in the Orbitrap with preview mode and monoisotopic precursor selection (MiPS) enabled. The top 10 precursors ions based on intensity were fragmented in the ion trap using 35% normalized collision energy. Dynamic exclusion was enabled for 180 s with a repeat count of 1 for a 30 s duration, a list size of 500, and an exclusion mass tolerance of 10 ppm.

**PBMC.** Human peripheral blood mononuclear cells (PBMC) were obtained from fresh venous blood using a Ficoll gradient protocol (Hoek et al., manuscript in preparation). PBMCs were lysed in 50% trifluoroethanol (TFE) in 50 mM triethylammonium bicarbonate (TEAB) essentially as described.[21] The total protein content was quantified using a Bradford assay. The PBMC lysate was reduced with DTT and alkylated with IAA. The sample was diluted 1:5 with 50 mM TEAB to make the final concentration of TFE <10% and digested with trypsin as previously described.[19] The desalted sample was analyzed using a LTQ-Orbitrap XL and a LTQ-Orbitrap Velos (Thermofisher). In the LTQ-Orbitrap XL analysis, 6-step MuDPIT experiments were performed using either MiPS or MiPS-off. A 100-$\mu$M i.d. fused silica microcapillary packed with 3 cm of Partisphere SCX material was coupled to a 100-$\mu$m i.d. fritless fused silica microcapillary column packed with 12 cm of Jupiter C18 RP material. For both the MiPS and MiPS-off experiments, the MudPIT salt steps were 50 mM, 100 mM, 200 mM, 300 mM, 500 mM, and 1 M ammonium acetate. A 100 min RP linear gradient from 5% to 45% Buffer B was used for each step. Buffer A was 0.1% formic acid in HPLC-grade water, and Buffer B was 0.1% formic acid in acetonitrile. A precursor ion scan was performed in the Orbitrap with preview mode enabled. The top 10 precursors ions based on intensity were fragmented in the ion trap using 35% normalized collision energy. Dynamic exclusion was enabled for 180 s with a repeat count of 1 for a 30 s duration and a list size of 500. The exclusion mass tolerance was set to 10 ppm for MiPS. For MiPS-off, the exclusion mass width was set between 1.5 and 2.5. For the LTQ-Orbitrap Velos analysis, 11-Step MuDPIT experiments were performed with either MiPS or MiPS-off, similar to Orbitrap XL experiments. The MudPIT salt pulses used were 25 mM, 50 mM, 75 mM, 100 mM, 150 mM, 200 mM, 250 mM, 300 mM, 500 mM, 750 mM, and 1 M ammonium acetate. A 100-$\mu$M i.d. fused silica microcapillary packed with 3 cm of PartiSphere SCX material was coupled with a fritless, microcapillary 100-$\mu$m i.d. column packed with 20 cm of Jupiter C18 RP material. A 90 min linear gradient from 2% to 40% Buffer B was used for each salt step. Buffer A was 0.1% formic acid in HPLC-grade water, and Buffer B was 0.1% formic acid in acetonitrile. A precursor ion scan was performed in the Orbitrap with preview mode enabled. The top 16 precursor ions based on intensity were fragmented in the dual pressure ion trap with 35% normalized collision energy. For both MiPS and MiPS-off, dynamic exclusion was enabled for 15 s with a repeat count of 1 set for a 10 s duration, a list size of 500, and an exclusion mass width set between 0.5 and 0.5.

**PMN.** Polymorphonuclear cells (PMN) were obtained from fresh venous human blood using a Ficoll gradient protocol (Hoek et al., manuscript in preparation). PMNs were lysed in 50% TFE in 50 mM HEPES buffer essentially as described.[21] The total protein content was quantified using a BCA assay. The PMN lysate was reduced with DTT and alkylated with IAA. The sample was diluted 1:5 with 50 mM TEAB to make the final concentration of TFE <10% and digested with trypsin as previously described.[19] A 6-step MuDPIT experiments was performed on a LTQ mass spectrometer (ThermoFisher) as previously described.[19]

### Tandem MS Data Analysis with SEQUEST

The RAW files generated from the LCQ, LTQ, and Orbitrap LC−MS/MS experiments were converted to dat or mzXML formats with the program ReadW. The MS/MS spectra were extracted from the mzXML file using the program MzXML2Search, and the data were analyzed using the SEQUEST algorithm to search either a Sigma48, *S. cerevisiae* (SGD_2010), or human Uniprot (uni280910) target and decoy concatenated protein database.[22,23] All decoy databases were created by reversing the sequences in the target databases. For Percolator, separate target and decoy searches were performed. For all data processing, a static modification of 57.021464 for cysteine was used. All SEQUEST searches were performed with no enzyme specificity.

### Analysis of SEQUEST Database Search Result Using PeptideProphet

To validate the PSMs identified by SEQUEST, the SEQUEST outputs from the LC−MS/MS experiments were loaded into the Trans-Proteomic Pipeline V.4.0.2 (TPP). The search outputs were converted to pep.XML format files and analyzed by the TPP program PeptideProphet.[24] Validation of the PSMs was performed by testing a range of probability filters until the desired FDR was reached. The pep.XML output file from PeptideProphet was converted to a CSV format. The CSV file was parsed with the in-house Perl script digest4peptide.pl, to sort the validated PSMs into lists of fully, half-, or non-canonical tryptic peptide consensus sequences.

### Analysis of SEQUEST Database Search Using ATP

The SEQUEST *.out files were concatenated by an in-house Perl script grab_files_threaded.pl to generate a merged *.outs file. The concatenated *.outs file was parsed and loaded into an Oracle relational database using the in-house Perl script concurrent_loading.pl and processed and analyzed using BIGCAT/ATP.[25,26] Two previously described filters, with low and high thresholds, were used to validate PSMs.[10,27] The low-threshold filter for PSMs was set with cutoff values of $Xcorr \geq 1.5$ for +1 charge state spectra, $Xcorr \geq 2$ for +2 spectra, and $Xcorr \geq 2$ for +3 spectra. Only fully canonical PSMs were accepted.[10] For a high-threshold filter,[27] PSMs with a +1 charge state were valid if they were fully canonical and had an $Xcorr > 1.9$. PSMs with a +2 charge state were valid if they were fully caonical or half-canonical and had $Xcorr$ ranges between 2.2 and 3.0. PSMs with a +2 charge state and an $Xcorr$ >3.0 were valid regardless of the PSM's protease consensus pattern. Finally, +3 peptides were valid if they were fully or half-canonical and had an $Xcorr$ >3.75. The filtered outputs from both filters were stored in CSV-formatted files and analyzed using Microsoft Excel.

Table 1. Summary of LC−MS/MS Data Set and SEQUEST Search Result

| sample | mass spectrometer | MiPS | PSM[a] | target PSM | decoy PSM | FDR | decoy PSMs/total PSMs |
|---|---|---|---|---|---|---|---|
| Gcn4 | LCQ | N/A | 14892 | 6703 | 8189 | 1.09 | 0.55 |
| UPS1 | LTQ | N/A | 17335 | 8974 | 8361 | 0.96 | 0.48 |
| Tal08 | Orbitrap XL | ON | 69560 | 42222 | 27338 | 0.79 | 0.39 |
| PBMC | Orbitrap XL | ON | 103679 | 68334 | 35345 | 0.64 | 0.34 |
| PBMC | Orbitrap XL | OFF | 117751 | 76395 | 41356 | 0.70 | 0.35 |
| PBMC | Orbitrap Velos | ON | 301879 | 208765 | 93114 | 0.62 | 0.31 |
| PBMC | Orbitrap Velos | OFF | 447350 | 307549 | 139801 | 0.62 | 0.31 |
| PMN | LTQ | N/A | 51423 | 29992 | 21431 | 0.83 | 0.42 |

[a]Peptide spectrum match.

### Analysis of SEQUEST Database Search Using Percolator

The target and decoy SEQUEST outputs from the LC−MS/MS experiments were converted to a merged file in SQT format[28] using an in-house modified version of the program Unitemare.pl (http://fields.scripps.edu/downloads.php). The UNIX *sed* utility was used to remove the header information of the converted SQT files. Two entries, H SQT Generator SEQUEST and H SQTGeneratorVersion2.7, were added as headers to the SQT files so that they could be analyzed by Percolator. The SEQUEST target and decoy search results in SQT format were loaded into Percolator. A range of *q*-values were tested until the desired FDR was reached. The outputs were stored in tab delimited format. The outputs were parsed by the in-house Perl script get_digest4percolator.pl to sort the validated PSMs based into list of fully, half-, or non-canonical tryptic peptide consensus sequences.

### Analysis of SEQUEST Database Search Using the De-Noise Algorithm

The SEQUEST result in *.out format was converted to Microsoft Excel format and processed by the De-Noise algorithm implemented with Matlab version 7.8.0.347 running on a Dell T410 with the Windows Server 2008R2 Standard operating system, 32 GB of RAM, and an Intel Xeon CPU at 2.27 GHz. Two support libraries and packages LibSVM library[29] and SKMsmo[30] needed for De-Noise decision function calculation were installed on the Dell T410.[30] The Matlab functions tic and toc were used to measure the De-Noise execution time running on the T410 machine. The validated peptides generated by the De-Noise algorithm were exported in Microsoft Excel format.

## ■ RESULTS AND DISCUSSION

### Generation of Test Data Sets

One goal in developing the De-Noise algorithm was to create a PSM validation tool unbiased in terms of sample, type of mass spectrometer, or mass spectrometry method. Therefore we used seven LC−MS/MS data sets generated from a variety of control and experimental protein samples analyzed on different mass spectrometers. First, we used a prepared mixture of 48 known human proteins (UPS1), which allowed us to unambiguously identify incorrect and correct PSMs. Second, we used affinity purified yeast Gcn4 and Tal08 minichromosome complexes and Ficoll-purified human PBMCs, which were authentic biological samples and contained both expected and unexpected proteins. The Tal08 sample contained bovine serum album as an internal standard. All samples were trypsin-digested prior to LC−MS/MS analysis.

The Gcn4 and UPS1 samples were analyzed on LCQ and LTQ mass spectrometers, respectively. To maximize the number of precursor ions fragmented, these ion trap data sets measured the precursor ion masses at low resolution and mass accuracy.[31] The Tal08 and PBMC samples were analyzed on LTQ-Orbitrap instruments. The Orbitrap's high resolution and mass accuracy measurement of the precursor ions' $m/z$ ratios allowed us to use monoisotopic precursor selection (MiPS) to select only peptide-like precursors for fragmentation.[32−34] For the seven experiments, we searched the LC−MS/MS data against concatenated target and decoy databases using the SEQUEST algorithm with no protease enzyme specificity.[9] The top scoring SEQUEST PSM for each MS/MS spectrum was used for all downstream data validation. For spectra matched to peptide sequences in the target database, we assumed they were true hits and treated them as correct PSMs. Spectra matched to peptide sequences in the decoy database were considered false hits and were treated as incorrect PSMs. Based on this assumption, the FDR was computed using the following equation[35,36]

$$\mathrm{FDR} = 2D/(D + T)$$

where $D$ is the number of the spectra matched to decoy peptide sequences and $T$ is the total number of the PSMs matched to target peptide sequence. Table 1 summarizes the different samples, types of mass spectrometers, precursor ion selection methods, the number of MS/MS spectra collected, and the unfiltered SEQUEST search results using concatenated target and decoy protein databases. These data sets and PSMs were used for the design and evaluation of the De-Noise algorithm.

### Variation in the Data Sets

We observed several effects of the type of MS/MS analysis on the data sets. First, we compared the data sets obtained from the LCQ and LTQ to those obtained from the LTQ-Orbitrap XL and LTQ-Orbitrap Velos. There were distinct differences in the FDRs and decoy PSMs/total PSMs ratios when we compared the UPS1 and Gcn4 results to the Tal08 and PBMC results (Table 1). The LCQ and LTQ data had higher calculated FDRs and decoy PSM/total PSMs ratios compared the Orbitrap data. Compared to the LCQ and LTQ instruments, the Orbitrap mass spectrometers have higher mass accuracy and resolution and different precursor methods for selecting and analyzing precursor ions. Second, we found that the precursor selection feature (MiPS) of the Orbitrap instruments influenced the results. With MiPS, there were significantly fewer PSMs, target PSMs, and decoy PSMs compared to with MiPS-off (Table 1). Although the precursor selection method significantly influenced the total number of PSMs, there was little variation in the decoy PSMs/total PSMs ratios and the calculated FDRs if MiPS or MiPS-off was used. We postulated that these variations in the quality of the acquired mass spectrometry data needed to be taken into consideration in the design of the De-Noise algorithm.

### Creating an Improved Data Set for the Decision Function

In a typical binary classification of correct and incorrect PSMs, the target PSMs are labeled as correct or +1, and decoy PSMs are labeled as incorrect or −1. The classifier learns from the training data set to assign either +1 or −1 class labels to PSMs. However, a large number of PSMs assigned as target PSMs by SEQUEST are actually incorrect.[2] These mislabeled target PSMs should be assigned to the incorrect class. If the mislabeled target PSMs were discarded from the PSMs data set, we would create a better target PSM data set to generate an improved decision function. A challenge was how to eliminate the incorrect target PSMs from the correct target PSMs.

The initial step in De-Noise is to cleanse the data set of incorrect or noisy target PSMs. The PSM data were represented as vectors based on the attributes from five SEQUEST scores: *Xcorr*, *DeltaCn*, *SPrank*, *Ions*, and *Calc-neutral-pep-mass*. To avoid attributes with larger values dominating ones with smaller values, we normalized each of the original SEQUEST scores by using the equation

$$x\_nor = x\_raw - (\text{mean of } x\_raw)/(std x\_raw)$$

where $x\_nor$ is the normalized SEQUEST score, $x\_raw$ is the original SEQUEST score, and $std\ x\_raw$ is the standard deviation of the original SEQUEST score. Next, each target PSMs is classified as incorrect or correct by computing its distance to the centroid of the decoy PSMs that belongs to the −1 class. Specifically, given a set $S$ with $m$ PSMs, $S$ consisted of $m_+$ +1 and $m_-$ −1 PSMs. Let $S^+ = [x^+_1, x^+_2, ..., x^+_{m+}]$ be the set of +1 PSMs and $S^- = [x^-_1, x^-_2, ..., x^-_{m-}]$ be the set of −1 PSMs. We first apply multiple kernel methods to map the PSM data points using the function $\phi$ into feature space where the target and decoy can be separated more easily.[37] The centroid of the −1 PSMs in the feature space denoted by $x^-_C$ is computed using the equation

$$x^-_C = \frac{1}{m_-}\sum_{i=1}^{m_-}\phi(x_i)$$

where $\phi$ is the mapping function mapping PSM data points to feature space in which the inner product $\langle\phi(x), \phi(y)\rangle$ can be computed through the kernel function.[37] Next, we computed the distance between each target PSM data point $x_i$ and $x^-_C$ in the feature space as follows

$$\begin{aligned} d(x_j, x^-_C) &= \|\phi(x_j) - \frac{1}{m_-}\sum_{i=1}^{m_-}\phi(x_i)\| \\ &= \langle\phi(x_j) - \frac{1}{m_-}\sum_{i=1}^{m_-}\phi(x_i), \phi(x_j) - \frac{1}{m_-}\sum_{i=1}^{m_-}\phi(x_i)\rangle^{1/2} \\ &= (k(x_j, x_j) - \frac{2}{m_-}\sum_{i=1}^{m_-}k(x_j, x_i) + \frac{1}{m_-^2}\sum_{i=1}^{m_-}\sum_{k=1}^{m_-}k(x_i, x_k))^{1/2} \end{aligned}$$

where $k(x_i, x_j)$ is the kernel function.

All +1 target PSMs with distances to the centroid of the −1 decoy PSMs less than a specific threshold are assumed to be incorrect and are discarded from the target PSM data set. Let $d(x^+_i, x^-_C)$ be the distance from $x^+_i$ to $x^-_C$ in the feature space. For a threshold $d_0$, $x^+_i$ were selected as incorrect and should be removed if $d(x^+_i, x^-_C) \le d_0$. In practice, it was a challenge to choose the optimal threshold $d_0$ because correct target PSMs could be discarded if $d_0$ is set too stringently. If it is too permissive, a large number of incorrect PSMs would remain in the data set.

Instead of using a distance threshold, we set the number of discarded incorrect target PSMs by applying the ratio $\theta$ using the equation below.

$$\theta = \text{assumed false positive PSMs/total PSMs}$$

Previous studies have shown 10−50% of target PSMs identified by a database search engine are correct.[2] The dilemma was how to determine $\theta$ to generate a target PSMs data set cleansed of incorrect PSMs without throwing out correct target PSMs.

To guide our determination of $\theta$, we took advantage of the distinct differences in the decoy PSMs/total PSMs ratio between the unfiltered LCQ and LTQ data sets and those derived from the Orbitrap data sets (Table 1). We assumed that the precursor ion spectra collected using the Orbitrap are of higher quality compared to the data sets obtained using LCQ and LTQ ion traps. A majority of MS/MS spectra (>60%) collected from the Orbitrap instruments were assigned to target PSMs, whereas only ~50% of the spectra collected from the LCQ and LTQ were assigned to target PSMs. We therefore inferred that if the decoy PSMs/total PSMs ratio is ≥40%, $\theta$ needs to be set to a larger value to cleanse more incorrect target PSMs. However, if the decoy PSMs/total PSMs ratio was <40%, we needed to set $\theta$ at a smaller value. On the basis of previous studies, $\theta$ was set at 0.1 for the UPS1 and Gcn4 data sets.[2] This resulted in the elimination of additional incorrect target PSMs. For the Tal08 and PBMC data sets, we empirically optimized the values for $\theta$ by testing a $\theta$ range of 0.01−0.1. We found a $\theta$ of 0.03 resulted in the optimal number of incorrect target PSMs being cleansed from the target PSMs while keeping the distribution of fully, half-, and non-canonical PSMs consistent with the expected ratios (Supplementary Figure S1).

The De-Noise algorithm was designed to automatically determine the value of $\theta$ based on the decoy PSMs/total PSMs ratio from the unfiltered SEQUEST results. With the selected $\theta$, the incorrect target PSMs were iteratively eliminated based on the distance of the target PSMs to the centroid of decoy PSMs. Target PSMs with the shortest distance to the centroid of the decoy PSMs are discarded first. This elimination of target PSMs considered noisy continued until the eliminated PSMs/total PSMs ratio satisfied $\theta$.

### Refining the PSM Data sets Using SVM-based Decision Functions

Finding an efficient decision function to calculate the decision score for each PSM was a critical step in the De-Noise algorithm. A kernel-based method provides a powerful learning tool for data sets with nonlinear structures and is adaptable to a variety of data types. The kernel method works by mapping data points within the vector space into a feature space where they can be easily separated. With the kernel method, we could combine different mappings by the sum of corresponding kernel matrices to provide complementary views of the data. In order to precisely characterize the relationships of each pair of PSMs, we tested the Polynomial, Gaussian, and Laplace kernel functions. Because it gave the greatest separation between target and decoy PSMs (Supplementary Figure S2), we selected Gaussian kernels computed with different weights using the equation

$$K = \sum_{i=1}^{m}\mu_i K^i$$

where $K$ is the combination of individual Gaussian kernels, $K^i$, $i = 1, ..., m$, and $\mu_i$ are the corresponding weights. In our experimental studies, the kernel width of the individual Gaussian

Table 2. Summary of Unfiltered, Categorized PSMs

| | target[a] | | | decoy[b] | | | decoy/target ratio | | |
|---|---|---|---|---|---|---|---|---|---|
| data set | fully tryptic | half-tryptic | non-tryptic | fully tryptic | half-tryptic | non-tryptic | fully tryptic | half-tryptic | non-tryptic |
| Gcn4LCQ | 1453 | 1210 | 4040 | 106 | 1465 | 6618 | 0.07 | 1.211 | 1.638 |
| UPS1LTQ | 645 | 2013 | 6316 | 236 | 2588 | 5537 | 0.36 | 1.286 | 0.877 |
| Tal08 OriXL MiPS | 14893 | 6809 | 20520 | 419 | 5877 | 21042 | 0.028 | 0.863 | 1.025 |
| PBMC OriXL MiPS | 26760 | 15647 | 25927 | 737 | 8583 | 26025 | 0.03 | 0.548 | 1.004 |
| PBMC OriXL MiPS-off | 28561 | 17490 | 30344 | 948 | 10333 | 30075 | 0.03 | 0.590 | 0.991 |
| PBMC OriVel MiPS | 110404 | 35915 | 62446 | 2520 | 24682 | 65912 | 0.023 | 0.687 | 1.056 |
| PBMC OriVel MiS-off | 134117 | 77052 | 96380 | 3414 | 34985 | 101402 | 0.025 | 0.454 | 1.052 |
| PMN LTQ | 8257 | 5946 | 15789 | 376 | 4752 | 16303 | 0.046 | 0.799 | 1.033 |

[a]Target PSMs. [b]Decoy PSMs.

kernel was chosen as 1, 0.5, and 0.2 respectively, and the weights $\mu_i$ were learned by using the SKMsmo software package.[30]

After the noisy target PSMs are discarded from the original target PSMs data set based on $\theta$, two rounds of data refining with SVM-based decision functions are performed to separate the correct from the incorrect PSMs. In the first round, the updated target PSMs data set $S^+$ and the decoy PSMs data set $S^-$ are combined into set $S_0$ to build the first SVM decision function. The target PSMs $x_i^+ \in S^+$ are treated as incorrect PSMs if their decision function $f(x_i^+) \leq 0$ where $f(x_i^+)$ was determined by the SVM learning model. These incorrect target PSMs are discarded from $S_0$ to generate $S_1$ containing the remaining PSMs.

We observed that a subset of $S_1$'s target PSMs had $f(x_i^+)$ scores < decoy PSMs $f(x_i^-)$ scores. We assumed some of these PSMs to be incorrect and targeted them in a second SVM decision function and refinement. The refined target PSM data set $S_1$ is used to build a second decision function. To remove the incorrect target PSMs, we applied a new parameter $\gamma$ using the equation

$$\gamma = \text{no. of retained decoy PSMs in } S_1 \,/\, \text{total no. of decoy PSMs in } S_1$$

In this second round, $f(x_\gamma^-)$ is the score of the lowest scoring decoy PSM that was retained. By comparing $f(x_i^+)$ to $f(x_\gamma^-)$, the target PSM $x_i^+$ is discarded if $f(x_i^+) \leq f(x_\gamma^-)$. De-Noise iteratively tests a range of $\gamma$ until the desired FDR is reached. As a result, a second set of incorrect target PSMs are cleansed to generate $S_2$. In the refining processes, a default slack penalty parameter of 1 is used.

## Refining Target PSMs Using Proteolytic Peptide Patterns

After the refining steps, De-Noise used proteolytic patterns to validate the PSMs from $S_2$. We catagorized PSMs in the seven data sets into three groups based on their protease digestion patterns: fully canonical, half-canonical, and non-canonical. Table 2 shows the majority of the PSMs representing fully canonical peptides were assigned to target peptides. Only a very small number of fully canonical PSMs were assigned to decoy peptides. These results indicated that the fully canonical target PSMs are more likely to be correct matches and should be retained. However, the decoy/target ratio showed the number of half- and non-canonical target and decoy PSMs assigned to the LCQ and LTQ data sets were almost equal (Table 2). This strongly implied that a higher percentage of false positives were present in half- and non-canonical target PSMs compared to the fully canonical PSMs. From this observation, we reasoned that a higher proportion of incorrect target PSMs from these two categories needed to be ultimately eliminated. We observed the Orbitrap data sets had similar decoy/target ratios for fully canonical and non-canonical PSMs. However, the Orbitrap data sets had a lower decoy/target ratio for the half-canonical target PSMs compared to the half-canonical decoy PSMs from the LCQ and LTQ data sets. We reasoned that a higher percentage of the half-canonical target PSMs from the Orbitrap data sets should be retained compared to data from the LCQ and LTQ.

In a proteolytic digest using a site-specific protease, a majority of peptides are canonical followed by half canonical and non-canonical peptides. Therefore, a greater weight is applied to canonical peptides compared to half-canonical and non-canonical peptides. Using De-Noise, all fully canonical peptides generated after the second refining were retained in the LCQ, LTQ, and Orbitrap data sets. However, the distribution of half- and non-canonical PSMs coming from De-Noise were significantly different compared to the distributions generated by PeptideProphet and Percolator. To correct for this result, the half- and non-canonical PSMs from the second round of SVM-based refining were filtered depending on the type of mass spectrometer used. We developed an approach that aims to remove half- and non-canonical PSMs after the second refining while optimizing both the number of validated PSMs and the distribution of half- and non-canonical PSMs. First, for the PSMs in data set $S_2$, we calculated a $\text{PSM}_{\text{evaluator}}$ value using the equation

$$\text{PSM}_{\text{evaluator}} = Xcorr_{\text{Nor}} + DeltaCn_{\text{Nor}}$$

where $Xcorr_{\text{Nor}}$ is the normalized *Xcorr* and $DeltaCn_{\text{Nor}}$ is the normalized *DeltaCn* described earlier. We used these two attributes because they have been previously shown to contribute the most toward measuring the accuracy and uniqueness of a PSM, respectively.[9,11,15] Next we calculated an Expectation factor ($\tau$) using the two equations

$$\tau_{\text{half}} = \text{no. of half-canonical PSMs accepted} \,/\, \text{total no. of half-canonical PSMs in } S_2$$

$$\tau_{\text{non}} = \text{no. of non-canonical PSMs accepted} \,/\, \text{total no. of non-canonical PSMs in } S_2$$

Using a range of $\tau_{\text{half}}$ and $\tau_{\text{non}}$, we developed an iterative approach to select the $\tau$ values that optimized the balance between the total number of validated PSMs and the distribution of half- and non-canonical PSMs. We tested a range of $\tau_{\text{half}}$ and $\tau_{\text{non}}$ to achieve an equivalent distribution of half- and non-canonical validated PSMs as reported by PeptideProphet and Percolator (Supplementary Tables S1 and S2). The values for $\tau_{\text{half}}$ and $\tau_{\text{non}}$ for the LCQ/LTQ data sets were determined to be 0.12 and 0.005, respectively (Supplementary Table S1). For Orbitrap data sets, $\tau_{\text{half}}$ and $\tau_{\text{non}}$

were determined to be 0.5 and 0.005 (Supplementary Table S2), respectively. With the $\tau_{half}$ and $\tau_{non}$ values, the half and non-canonical PSMs were discarded from the second round of refining based on the $PSM_{evaluator}$ score. Finally, the pseudocode for the entire De-Noise algorithm is summarized in Figure 1.

**Data:** $S^+, S^-, \theta, \gamma$
**Result:** $S^+$
Pre-process;
  Compute $x_{\bar{C}}$ and $d(x_i^+, x_{\bar{C}})$ for $i = 1, 2, \ldots, m$;
  Sort $d(x_i^+, x_{\bar{C}})$ for $i = 1, 2, \ldots, m$;
  Calculate the ratio of decoy PSMs to the total, and choose $\theta$ according to the ratio;
  Select $\theta|S^+|$ data points $x_i^+$ with smallest distance and remove them;
  Update $S^+$;
  Set $S = S^+ \cup S^-$;
Refine process;
  Classification process;
    Train a SVM classifier $f$ based on $S$ ;
    Remove $x_i^+$ from $S^+$ if $f(x_i^+) \leq 0$;
    Update $S^+$;
    Set $S = S^+ \cup S^-$;
  Adjust process;
    Train a SVM classifier $f$ based on $S$;
    Sort $f(x_i^-)$ in descending order;
    Let $x_\gamma^-$ be the $\gamma|S^-|$th largest $f(x_i^-)$;
    Remove $x_i^+$ from $S^+$ if $f(x_i^+) \leq f(x_\gamma^-)$;
Post-process;
  Keep all full digested PSMs ;
  Keep the top $\tau_{half}$ half digested PSMs according to PSMevaluator;
  Keep the top $\tau_{non}$ non-digested PSMs according to PSMevaluator;

**Notation Summary**

| | |
|---|---|
| $x_i^+$ | the $i$th +1 PSM (target PSM) |
| $x_i^-$ | the $i$th −1 PSM (decoy PSM) |
| $x_{\bar{C}}$ | the centroid of all −1 PSMs in the feature space |
| $d(x_i^+, x_{\bar{C}})$ | the distance between $x_i^+$ and $x_{\bar{C}}$ in the feature space |
| $\theta$ | the ratio of assumed false positive PSMs to total PSMs |
| $m$ | total number of PSMs |
| $S$ | the set of $m$ PSMs |
| $S^+$ | the set of +1 PSMs |
| $S^-$ | the set of −1 PSMs |
| $S_0$ | the union of $S^-$ and $S^+$ updated after the first round |
| $S_1$ | the set of PSMs obtained by removing $x_i^+$ from $S_0$ if $f(x_i^+) \leq 0$ |
| $\gamma$ | the ratio of number of retained decoy PSMs in $S_1$ to the total number of decoy PSMs in $S_1$ |
| $S_2$ | the set of PSMs obtained by removing removing $x_i$ from $S_1$ if $f(x_i) \leq f(x_\gamma^-)$ |
| $\tau_{half}$ | the ratio of the number of half-canonical PSMs accepted to the total number of half-canonical PSMs in $S_2$ |
| $\tau_{non}$ | the ratio of the number of non-canonical PSMs accepted to the total number of non-canonical PSMs in $S_2$ |

**Figure 1.** Pseudocode for the De-Noise algorithm.

## Evaluating De-Noise's Performance

To evaluate the performance of De-Noise, we compared its performance against the PeptideProphet and Percolator algorithms for validating SEQUEST target PSMs.[10,11] First, we measured De-Noise's runtime using the UPS1 and PBMC Oritrap Velos MiPS-off data sets. It took ∼43 s for De-Noise to process the UPS1 data, which was the smallest data set, and ∼4082 s to validate the PBMC Oribitrap Velos MiPS-off data, which was the largest. We found the runtimes were very similar to those of comparable approaches.[13−15] Second, we compared the number of PSMs identified by three algorithms. Table 3 shows De-noise validated more PSMs than the semi-supervised learning approaches PeptideProphet and Percolator and the two Xcorr filtering approaches (Low- and High-Stringency). Since the approaches using learning algorithms to validate PSMs from SEQUEST had the highest performance (Table 3), we focused our evaluation of De-Noise compared to PeptideProphet and Percolator.

To compare the validated PSMs from the De-Noise, PeptideProphet, and Percolator, we looked at the overlapping PSMs. Figure 2 and Supplementary Figure S3 show that the majority of De-Noise validated PSMs were also validated by PeptideProphet and Percolator. Table 4 shows a numerical summary of the overlapping validated PSMs from the three approaches. For example, for UPS1 data set, 94% of the PSMs validated by PeptideProphet overlapped with De-Noise, while 90% of the PSMs validated by Percolator were validated by De-Noise. Similar patterns were seen in the other data sets in which De-Noise shared more validated PSMs with PeptideProphet than with the Percolator (Table 4, Figure 2, and Supplementary Figure S3).

To show that our approach to remove half- and non-canonical target PSMs after the refining generated a similar distribution compared to PeptideProphet and Percolator, we compared the categorized overlapping outputs from the three approaches for the UPS1, Gcn4, and Tal08 data sets (Tables 5 and 6). The validated PSMs and the overlapped PSMs from all three approaches for UPS1 and Gcn4 showed a similar distribution pattern. The number of validated fully canonical PSMs was the largest class followed by the half-canonical and non-canonical PSMs, respectively (Tables 5 and 6). The considerable overlap of validated PSMs from De-Noise, PeptideProphet, and Percolator (Table 5) and the similar distributions of PSMs (Table 6) showed De-Noise's approach to retain the most significant half- and non-canonical PSMs was valid.

We compared De-Noise, PeptideProphet, and Percolator using different FDR values. Figure 3 shows the number of

**Table 3. Summary of PSMs Validated by Different Approaches**

| | Gcn4 LCQ | | UPS1 LTQ | | Tal08 OriXL MiPS | | PMN LTQ | |
|---|---|---|---|---|---|---|---|---|
| approach | target PSMs | decoy PSMs | target PSMs | decoy PSMs | target PSMs | decoy PSMs | target PSMs | decoy PSMs |
| SEQUEST | 6703 | 8189 | 8974 | 8361 | 42222 | 27338 | 29992 | 21431 |
| PeptideProphet | 1443 | 38 | 566 | 18 | 15638 | 387 | 8957 | 220 |
| Percolator | 1394 | 35 | 438 | 12 | 14371 | 354 | 9060 | 219 |
| De-Noise | 1488 | 39 | 811 | 21 | 18454 | 475 | 11341 | 287 |
| Low-Stringency | 1128 | 18 | 293 | 6 | 9979 | 108 | 11252 | 553 |
| High-Stringency | 588 | 11 | 154 | 0 | 7083 | 272 | 8411 | 134 |
| | PBMC OriXL MiPS | | PBMC OriXL MiPS-off | | PBMC OriVel MiPS | | PBMC OriVel MiPS | |
| approach | target PSMs | decoy PSMs | target PSMs | decoy PSMs | target PSMs | decoy PSMs | target PSMs | decoy PSMs |
| SEQUEST | 68334 | 33545 | 76395 | 41356 | 208372 | 93112 | 307459 | 139081 |
| PeptideProphet | 33233 | 802 | 35673 | 869 | 120961 | 2947 | 175790 | 4393 |
| Percolator | 33053 | 793 | 35230 | 866 | 122568 | 3133 | 173719 | 4363 |
| De-Noise | 35070 | 813 | 37894 | 927 | 128977 | 3272 | 176206 | 4287 |
| Low-Stringency | 6135 | 235 | 25761 | 346 | 107693 | 1372 | 127857 | 1728 |
| High-Stringency | 11768 | 335 | 20963 | 930 | 71876 | 6539 | 120180 | 9043 |


dx.doi.org/10.1021/pr300631t | J. Proteome Res. 2013, 12, 1108−1119

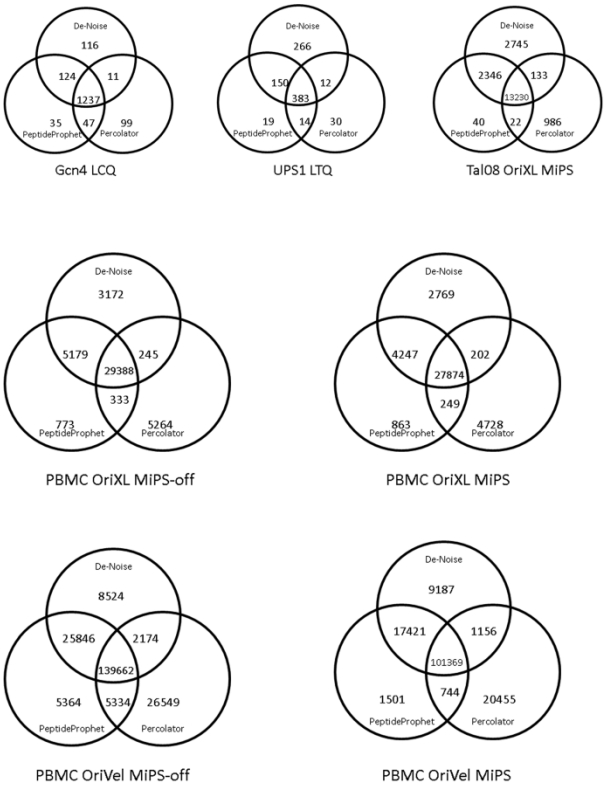



**Figure 2.** Venn diagrams for the seven data sets showing the number of overlapping validated PSMs from De-Noise, PeptideProphet, and Percolator. An FDR of 0.05 was used for all three approaches.

**Table 4. Summary of Overlapping, Validated PSMs[a]**

| data set | PSMs shared between PeptideProphet and De-Noise | % PeptideProphet shared by De-Noise | PSMs shared between Percolator and De-Noise | % Percolator Shared by De-Noise | PSMs shared between Percolator and PeptideProphet | % Percolator Shared by PeptideProphet |
|---|---|---|---|---|---|---|
| Gcn4 LCQ | 1361 | 94 | 1248 | 90 | 1284 | 92 |
| UPS1 LTQ | 533 | 94 | 395 | 90 | 397 | 90 |
| Tal08 OriXL-MiPS | 15576 | 99 | 12763 | 89 | 13252 | 92 |
| PBMC OriXL-MiPS | 32121 | 97 | 28076 | 85 | 28123 | 85 |
| PBMC OriXL MiPS-off | 34567 | 97 | 29633 | 84 | 29721 | 84 |
| PBMC OriVel MiPS | 118790 | 98 | 102525 | 84 | 102113 | 83 |
| PBMC OriVel MiPS-off | 165508 | 94 | 141836 | 82 | 144996 | 83 |

[a]Data used in this table are outputs under FDR = 0.05.

validated PSMs from the seven data sets using a series of FDRs. The performance of a validation approach is better if it validates more target PSMs compared to another approach with the same FDR. The plots in Figure 3A and B demonstrates that De-Noise validated more target PSMs compared to PeptideProphet and Percolator. From the Gcn4 data set, De-Noise validated approximately 2.8% and 13.2% more PSMs than PeptideProphet and Percolator, respectively. Likewise, from the UPS1 data set, De-Noise identified about 31% and 73% more PSMs than PeptideProphet and Percolator, respectively. For the Tal08 data set, the validated PSMs increased 30% and 29% using De-Noise compared to PeptideProphet and Percolator, respectively. A similar pattern is seen in PBMC data sets (Figure 3D–G). We observed that De-Noise consistently outperforms Peptide Prophet and Percolator in terms of the number of target PSMs validated at a given FDR.

## Evaluating De-Noise for Sensitivity and Specificity

From an applied mathematics point of view, distinguishing correct from incorrect PSMs can be treated as a two-class classification

Table 5. Summary of Overlapping Validated PSMs[a]

| | PSMs shared between PeptideProphet and De-Noise | | | PSMs shared between Percolator and De-Noise | | | PSMs shared between PeptideProphet and Percolator | | |
|---|---|---|---|---|---|---|---|---|---|
| data set | full | half | non | full | half | non | full | half | non |
| Gcn4 LCQ | 1314 | 47 | 0 | 1206 | 42 | 0 | 1237 | 46 | 1 |
| UPS1 LTQ | 399 | 125 | 9 | 276 | 112 | 7 | 270 | 115 | 12 |
| Tal08 OriXL MiPS | 14528 | 1041 | 7 | 12881 | 482 | 0 | 12767 | 485 | 0 |
| PBMC OriXL MiPS | 25962 | 6046 | 113 | 23565 | 4423 | 88 | 23431 | 4566 | 126 |
| PBMC OriXL MiPS-off | 27604 | 6802 | 161 | 24420 | 5070 | 143 | 24247 | 5204 | 270 |
| PBMC OriVel MiPS | 107595 | 11149 | 46 | 94718 | 7771 | 36 | 93830 | 8171 | 112 |
| PBMC OriVelMiPS-off | 130259 | 34738 | 511 | 115919 | 25735 | 182 | 114134 | 29234 | 1628 |

[a]Data used in this table are outputs under FDR = 0.05.

Table 6. Distribution of Validated PSMs

| | De-Noise | | | PeptideProphet | | | Percolator | | |
|---|---|---|---|---|---|---|---|---|---|
| data set | full | half | non | full | half | non | full | half | non |
| Gcn4 LCQ | 1370 | 106 | 12 | 1375 | 68 | 1 | 1342 | 51 | 1 |
| UPS1 LTQ | 597 | 190 | 24 | 403 | 147 | 16 | 278 | 144 | 17 |
| Tal08 OriXL MiPS | 14860 | 3471 | 123 | 14539 | 1088 | 11 | 13855 | 516 | 0 |
| PBMC OriXL MiPS | 26699 | 8228 | 165 | 25977 | 7006 | 250 | 27025 | 5865 | 163 |
| PBMC OriXL MiPS-off | 28498 | 9292 | 194 | 27622 | 7649 | 402 | 28271 | 6661 | 298 |
| PBMC OriVel MiPS | 110138 | 18490 | 349 | 107730 | 13001 | 230 | 11990 | 10453 | 125 |
| PBMC OriVelMiPS-off | 133846 | 41752 | 608 | 130321 | 42835 | 2633 | 133436 | 38069 | 2214 |

problem, in which a classifier labels the PSMs as either true or false.[4,38] There are four possible outcomes from a binary classifier. If the classifier validates a PSM as true and the spectrum is also matched to a target peptide sequence, then it is called a true positive (TP). However, if the classifier validates a PSM as true but the spectrum was matched to a decoy peptide sequence, then it is said to be a false positive (FP). Conversely, if the classifier assigns a PSM as false and the spectrum was matched to a decoy peptide sequence, then a true negative (TN) has occurred. Finally, if the classifier assigns a PSM as false and the spectrum was matched to a target sequence, a false negative (FN) has occurred. The true positive rate (TPR) is defined as the ratio of the target PSMs validated by a classifier to the total number of target PSMs:

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$$

The false positive rate (FPR) is the ratio of the number of decoy PSMs falsely identified as a target PSMs to the total number of decoy PSMs:

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$$

The performance of a binary classification method is measured using two statistical parameters: sensitivity and specificity. Sensitivity is equal to the TPR and reflects the classifier's capability to correctly validate target PSMs from a pool of target PSMs. Specificity is equal to 1 − FPR and measures the frequency at which the classifier correctly validates decoy PSMs from the total pool of decoy PSMs. The overall performance of the classifier can be represented with a receiver operating characteristic (ROC) curve,[39] which plots the true positive rate (sensitivity) versus the false positive rate (1 − specificity). Each point on the ROC curve represents sensitivity/specificity. When two classifiers are compared, the classifier with the higher sensitivity at a given specificity is considered the better classifier.

The performance of the De-Noise, PeptideProphet, and Percolator approaches was evaluated by using ROC plots[39] (Figure 4). In general, the classifier with ROC plot closest to the left-hand border is considered the most robust. The robustness of a classifier declines as its curve gets closer to the 45° diagonal of the ROC space. Another index to assess the robustness of a classifier from a ROC plot is the area under the curve (AUC). The larger the area covered, the more robust is the classifier. For example, we calculated the AUC for the UPS1 data set using the trapezoid rule for an FPR range from 0.0008 to 0.0074.[40] For this FPR range, the AUC calculated for the three approaches shows De-Noise is the more robust approach (De-Noise (0.0006), PeptideProphet (0.00038), and Percolator (0.00025 (Figure 4B). For the other six data sets, the AUC for De-Noise was consistently larger compared to PeptideProphet and Percolator, showing that De-Noise was more robust (Figure 4A–G). Using the ROC curves, we compared the sensitivity for the three approaches. For the seven data sets, De-Noise consistently had the higher TPR compared to PeptideProphet and Percolator in the FPR range 0.01–0.05 (Figure 4A–G).

### Evaluating De-Noise's Performance Using an Independent Data set

Finally, to test De-Noise's performance in validating PSMs on a data set not used in its optimization, the De-Noise algorithm was applied to a human PMNs extract run on a LTQ ion trap mass spectrometer. The results from the SEQUEST search for the PMN LTQ data set are shown in Tables 1 and 2. Table 3 shows De-Noise validated more target PSMs compared to other validation approaches.

In summary, we have developed a highly sensitive and specific algorithm to validate PSMs from the SEQUEST search engine. The novel De-Noise algorithm first uses a data cleaning step based on the distance of the target PSMs to the centroid of the decoy PSMs to remove noisy, incorrect PSMs from the target PSMs. Second, De-Noise performs two rounds of data refining using SVM-based decision functions to validate correct target PSMs. Finally, the algorithm uses proteolytic information and the quality of the mass spectrometry data to perform a final validation. Using a variety of data sets based on different samples, mass spectrometers, and popular validation approaches, we show the De-Noise algorithm has improved sensitivity and specificity

in the 1−5% FDR range that is commonly used to report the accepted peptide sequences from tandem mass spectrometry search engines.

## ■ ASSOCIATED CONTENT

### Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION


### Corresponding Author

*Tel: 615-343-6823. Fax: 615-343-7392. E-mail: andrew.link@vanderbilt.edu.

### Author Contributions

#These authors contributed equally to this work.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant GM064779 and Vanderbilt University School of Medicine IDEAS Program grant 1-04-066-9530. This project has been funded in part with federal funds from the National Institutes of Allergy and Infectious Disease, National Institutes of Health, Department of Health and Human Services, under Contract No. 272200800007C and the Vanderbilt Clinical and Translational Science Award grant NIH RR024975. The software in this manuscript is available upon request. We thank Drs. Tony Weil and Kristen Hoek for allowing us to cite unpublished results.

## ■ ABBREVIATIONS

LC−MS/MS, liquid chromatography coupled with tandem mass spectrometry; PSM, peptide spectra match; FDR, false discovery rate; TPR, true positive rate; FPR, false positive rate; ROC, receiver operating characteristic curve; TP, true positive; TN, true negative; FP, false positive; FN, false negative; SVM, support vector machines; UPS1, Universal Proteomics Standard Set; DTT, dithiothreitol; IAA, iodoacetamide; RP, reverse-phase; nanoESI, nanoelectrospray ionization; MS, mass spectrometry; MS/MS, tandem mass spectrometry; PBMC, peripheral blood mononuclear cells; PMN, polymorphonuclear cells; TEAB, triethylammonium bicarbonate; TFE, trifluoroethanol; MiPS, monoisotopic precursor selection; LC, liquid chromatography; SCX, strong cation exchange; TPP, Trans-Proteomics Pipeline; MudPIT, multidimensional protein identification technology; $D$, total number of spectra matched to decoy peptide sequences; $T$, total number of the PSMs matched to target peptide sequence; AUC, area under the curve; $m^+$, number of target PSMs; $m^-$, number of decoy PSMs; $S^+$, the set of target PSMs; $S^-$, the set of decoy PSMs

## ■ REFERENCES


(1) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005**, *2*, 667−675.

(2) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207−214.

(3) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2003**, *2*, 43−50.

(4) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7*, 29−34.

(5) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 47−50.

(6) Navarro, P.; Vazquez, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *J. Proteome Res.* **2009**, *8*, 1792−1796.

(7) Goloborodko, A. A.; Mayerhofer, C.; Zubarev, A. R.; Tarasova, I. A.; Gorshkov, A. V.; Zubarev, R. A.; Gorshkov, M. V. Empirical approach to false discovery rate estimation in shotgun proteomics. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 454−462.

(8) Lam, H.; Deutsch, E. W.; Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. Proteome Res.* **2010**, *9*, 605−610.

(9) Eng, J. K.; McCormack, A. L.; Yates, I.; J., R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(10) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17*, 676−682.

(11) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(12) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4*, 787−797.

(13) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(14) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 254−265.

(15) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.

(16) Andrews, S.; Tsochantaridis, I.; Hofmann, T. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*; MIT Press: Vancouver, British Columbia, 2002; pp 561−568.

(17) Bennett, K. P. Combining support vector and mathematical programming methods for classification. In *Advances in Kernal Methods: Support Vector Learning*; MIT Press: Cambridge, MA, 1999; pp 307−326.

(18) Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.* **2009**, *8*, 3737−3745.

(19) Sanders, S. L.; Jennings, J.; Canutescu, A.; Link, A. J.; Weil, P. A. Proteomics of the eukaryotic transcription machinery: Identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. *Mol. Cell. Biol.* **2002**, *22*, 4723−4738.

(20) Unnikrishnan, A.; Gafken, P. R.; Tsukiyama, T. Dynamic changes in histone acetylation regulate origins of DNA replication. *Nat. Struct. Mol. Biol.* **2010**, *17*, 430−437.

(21) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154−1169.

(22) Cherry, J. M.; Hong, E. L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **2012**, *40*, D700−705.

(23) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005**, *33*, D154−159.

(24) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150−1159.

in the 1−5% FDR range that is commonly used to report the accepted peptide sequences from tandem mass spectrometry search engines.

## ■ ASSOCIATED CONTENT

### Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION


### Corresponding Author

*Tel: 615-343-6823. Fax: 615-343-7392. E-mail: andrew.link@vanderbilt.edu.

### Author Contributions

#These authors contributed equally to this work.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant GM064779 and Vanderbilt University School of Medicine IDEAS Program grant 1-04-066-9530. This project has been funded in part with federal funds from the National Institutes of Allergy and Infectious Disease, National Institutes of Health, Department of Health and Human Services, under Contract No. 272200800007C and the Vanderbilt Clinical and Translational Science Award grant NIH RR024975. The software in this manuscript is available upon request. We thank Drs. Tony Weil and Kristen Hoek for allowing us to cite unpublished results.

## ■ ABBREVIATIONS

LC−MS/MS, liquid chromatography coupled with tandem mass spectrometry; PSM, peptide spectra match; FDR, false discovery rate; TPR, true positive rate; FPR, false positive rate; ROC, receiver operating characteristic curve; TP, true positive; TN, true negative; FP, false positive; FN, false negative; SVM, support vector machines; UPS1, Universal Proteomics Standard Set; DTT, dithiothreitol; IAA, iodoacetamide; RP, reverse-phase; nanoESI, nanoelectrospray ionization; MS, mass spectrometry; MS/MS, tandem mass spectrometry; PBMC, peripheral blood mononuclear cells; PMN, polymorphonuclear cells; TEAB, triethylammonium bicarbonate; TFE, trifluoroethanol; MiPS, monoisotopic precursor selection; LC, liquid chromatography; SCX, strong cation exchange; TPP, Trans-Proteomics Pipeline; MudPIT, multidimensional protein identification technology; $D$, total number of spectra matched to decoy peptide sequences; $T$, total number of the PSMs matched to target peptide sequence; AUC, area under the curve; $m^+$, number of target PSMs; $m^-$, number of decoy PSMs; $S^+$, the set of target PSMs; $S^-$, the set of decoy PSMs

## ■ REFERENCES


(1) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005**, *2*, 667−675.

(2) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207−214.

(3) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2003**, *2*, 43−50.

(4) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7*, 29−34.

(5) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 47−50.

(6) Navarro, P.; Vazquez, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *J. Proteome Res.* **2009**, *8*, 1792−1796.

(7) Goloborodko, A. A.; Mayerhofer, C.; Zubarev, A. R.; Tarasova, I. A.; Gorshkov, A. V.; Zubarev, R. A.; Gorshkov, M. V. Empirical approach to false discovery rate estimation in shotgun proteomics. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 454−462.

(8) Lam, H.; Deutsch, E. W.; Aebersold, R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J. Proteome Res.* **2010**, *9*, 605−610.

(9) Eng, J. K.; McCormack, A. L.; Yates, I.; J., R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(10) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17*, 676−682.

(11) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(12) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4*, 787−797.

(13) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(14) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 254−265.

(15) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.

(16) Andrews, S.; Tsochantaridis, I.; Hofmann, T. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*; MIT Press: Vancouver, British Columbia, 2002; pp 561−568.

(17) Bennett, K. P. Combining support vector and mathematical programming methods for classification. In *Advances in Kernal Methods: Support Vector Learning*; MIT Press: Cambridge, MA, 1999; pp 307−326.

(18) Spivak, M.; Weston, J.; Bottou, L.; Kall, L.; Noble, W. S. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.* **2009**, *8*, 3737−3745.

(19) Sanders, S. L.; Jennings, J.; Canutescu, A.; Link, A. J.; Weil, P. A. Proteomics of the eukaryotic transcription machinery: Identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. *Mol. Cell. Biol.* **2002**, *22*, 4723−4738.

(20) Unnikrishnan, A.; Gafken, P. R.; Tsukiyama, T. Dynamic changes in histone acetylation regulate origins of DNA replication. *Nat. Struct. Mol. Biol.* **2010**, *17*, 430−437.

(21) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154−1169.

(22) Cherry, J. M.; Hong, E. L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **2012**, *40*, D700−705.

(23) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005**, *33*, D154−159.

(24) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150−1159.

(25) McAfee, K. J.; Duncan, D. T.; Assink, M.; Link, A. J. Analyzing proteomes and protein function using graphical comparative analysis of tandem mass spectrometry results. *Mol Cell Proteomics* **2006**, *5*, 1497−1513.
(26) Niu, X.; McAfee, K. J.; Duncan, D. T.; Assink, M.; Link, A. J. A computational and analysis tool for proteomics research. *UT-ORNL-KBRIN Bioinformatics Summit 2008*; BMC Bioinformatics: Cadiz, KY, 2008; pp 22.
(27) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19*, 242−247.
(28) McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2162−2168.
(29) Chang, C. C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.
(30) Bach, F. R.; Lanckriet, G. R. G.; Jordan, M. I. *Fast Kernel Learning Using Sequential Minimal Optimization*; Computer Science Division, University of California: Berkeley, CA, 2004.
(31) Schwartz, J. C.; Senko, M. W.; Syka, J. E. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 659−669.
(32) Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; et al. The Orbitrap: a new mass spectrometer. *J Mass Spectrom* **2005**, *40*, 430−443.
(33) Makarov, A.; Denisov, E.; Kholomeev, A.; Balschun, W.; Lange, O.; et al. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **2006**, *78*, 2113−2120.
(34) Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.
(35) Jiang, X.; Han, G.; Ye, M.; Zou, H. Optimization of filtering criterion for SEQUEST database searching to improve proteome coverage in shotgun proteomics. *BMC Bioinf.* **2007**, *8*, 323.
(36) Jones, A. R.; Siepen, J. A.; Hubbard, S. J.; Paton, N. W. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **2009**, *9*, 1220−1229.
(37) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **1998**, *2*, 121−167.
(38) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2003**, *2*, 137−146.
(39) Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Lett.* **2006**, *27*, 861−874.
(40) Yeh, S.-T. Using trapezoidal rule for the area under a curve calculation. In *Proceedings of the Twenty-Seventh Annual SAS User Group International (SUGI) Conference*: Cary, NC, 2002; pp 229.

# Appendix AA – Manuscript – 6: The Yeast Eukaryotic Translation Initiation Factor 2B Translation Initiation Complex Interacts with the Fatty Acid Synthesis Enzyme YBR159W and Endoplasmic Reticulum Membranes

Molecular and Cellular Biology

# The Yeast Eukaryotic Translation Initiation Factor 2B Translation Initiation Complex Interacts with the Fatty Acid Synthesis Enzyme YBR159W and Endoplasmic Reticulum Membranes

**Christopher M. Browne, Parimal Samir, J. Scott Fites, Seth A. Villarreal and Andrew J. Link**
***Mol. Cell. Biol.*** **2013, 33(5):1041. DOI: 10.1128/MCB.00811-12.**
**Published Ahead of Print 21 December 2012.**

MCB
Journals.ASM.org

# The Yeast Eukaryotic Translation Initiation Factor 2B Translation Initiation Complex Interacts with the Fatty Acid Synthesis Enzyme YBR159W and Endoplasmic Reticulum Membranes

**Christopher M. Browne,[a] Parimal Samir,[a] J. Scott Fites,[b] Seth A. Villarreal,[b]* Andrew J. Link[c]**

Department of Biochemistry,[a] Interdisciplinary Graduate Program in the Biomedical and Biological Sciences,[b] and Department of Pathology, Microbiology, and Immunology,[c] Vanderbilt University School of Medicine, Nashville, Tennessee, USA

**Using affinity purifications coupled with mass spectrometry and yeast two-hybrid assays, we show the *Saccharomyces cerevisiae* translation initiation factor complex eukaryotic translation initiation factor 2B (eIF2B) and the very-long-chain fatty acid (VLCFA) synthesis keto-reductase enzyme YBR159W physically interact. The data show that the interaction is specifically between YBR159W and eIF2B and not between other members of the translation initiation or VLCFA pathways. A *ybr159wΔ* null strain has a slow-growth phenotype and a reduced translation rate but a normal GCN4 response to amino acid starvation. Although YBR159W localizes to the endoplasmic reticulum membrane, subcellular fractionation experiments show that a fraction of eIF2B cofractionates with lipid membranes in a YBR159W-independent manner. We show that a *ybr159wΔ* yeast strain and other strains with null mutations in the VLCFA pathway cause eIF2B to appear as numerous foci throughout the cytoplasm.**

In eukaryotic translation initiation, the eukaryotic translation initiation factor 2 (eIF2) bound with GTP is required to interact with the initiator $Met_i$-tRNA to form the ternary complex. After start codon recognition, eIF2-GTP is hydrolyzed to GDP, and eIF2-GDP dissociates from the translation initiation complex (1, 2). eIF2-GDP must exchange GDP with GTP before it can initiate another round of translation (Fig. 1A). The initiation factor eIF2B is an essential guanine nucleotide exchange factor (GEF) responsible for exchanging GDP for GTP on eIF2 (3). It is the only known target of eIF2B. This exchange reaction is one of the rate-limiting steps in translation initiation and is the target of numerous signaling pathways in yeast, as well as higher eukaryotes (4–10). Although the majority of eukaryotic GEFs are monomeric, eIF2B is unique among GEFs in that it is composed of multiple subunits. In *Saccharomyces cerevisiae*, eIF2B is composed of the five subunits GCD1, GCD2, GCN3, GCD6, and GCD7. The GCD6 subunit is necessary and sufficient for catalytic activity, although at a significantly reduced rate compared to the eIF2B complex (11–13). Coexpression of GCD6 with GCD1 yields similar GEF activity as the eIF2B holoenzyme (13). Of the other three subunits, previous studies show GCD2 and GCD7 to be involved in the stability of the complex and regulatory activity (13–15). GCN3 is required for eIF2B's role in the GCN4 stress response pathway (16, 17). With the exception of *GCN3*, all of the yeast eIF2B genes are essential (3).

Recent studies show that a significant fraction of yeast eIF2B resides in distinct foci in the cytoplasm known as "2B bodies" (18, 19). Green fluorescent protein (GFP) fluorescence microscopy shows the bodies contain both eIF2B and eIF2. The initiation factor eIF2 appears to shuttle in and out of the 2B bodies (18). The shuttling occurs quickly during logarithmic growth and slower following disruptions of translation initiation. The 2B bodies are thought to be the sites of eIF2B's GEF activity.

In eukaryotes, two distinct complexes are responsible for the synthesis of fatty acids (FA) (20, 21). The cytoplasmic fatty acid synthase (FAS) complex elongates fatty acids from 2 to 18 carbons in length in a four-reaction cycle. A second fatty acid synthesis complex, the elongase complex, is responsible for the elongation of fatty acids from 18 to 26 carbons (Fig. 1B) (22). The longer-chain fatty acids are known as very-long-chain fatty acids (VLCFAs). In *S. cerevisiae*, VLCFAs make up 1 to 5% of total fatty acids (22, 23) and the predominant VLCFA is 26 carbons long (24). The VLCFAs are crucial for the formation of lipid rafts in yeast (25). Although the FAS and elongase complexes share very similar catalytic steps, different sets of enzymes catalyze the elongation reactions in the two pathways (Fig. 1B). The elongase complex's enzymes localize to the endoplasmic reticulum (ER) membrane (26, 27). The complex receives fatty acids from the cytoplasmic FAS complex and elongates them to VLCFAs (28). Previous studies have shown that YBR159W, also known as IFA38, is a keto-acyl reductase required for the second step in the yeast elongase's pathway (Fig. 1B) (29, 30). A *ybr159wΔ* null yeast strain has a slow-growth phenotype and altered VLCFA lipid composition (30). Although both FEN1 and SUR4 catalyze the first enzymatic step in the elongase pathway, they are not redundant and are responsible for different-chain-length precursor fatty acids. FEN1 prefers 20-carbon-long precursors, while SUR4 has a broader range of chain length specificity but is required to convert 24-carbon-long VLCFAs to their final 26-carbon-long form (31). The elongase enzymes TSC13 and PHS1 are both essential (32, 33). In yeast, newly synthesized VLCFAs are predominantly incorporated first into ceramide and eventually into sphingolipids (24).

Received 14 June 2012 Returned for modification 12 July 2012
Accepted 16 December 2012
Published ahead of print 21 December 2012
Address correspondence to Andrew J. Link, andrew.link@vanderbilt.edu.
* Present address: Seth A. Villarreal, Case Western Reserve University School of Medicine, Department of Pharmacology and Cleveland Center for Membrane and Structural Biology, Cleveland, Ohio, USA.

doi:10.1128/MCB.00811-12

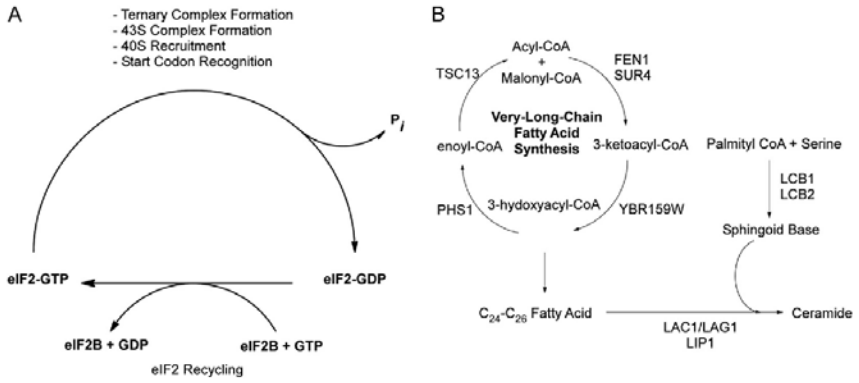



FIG 1 eIF2B and the VLCFA functional pathways. (A) Diagram showing the GEF pathway of eIF2B that is required for recharging eIF2 with GTP to begin a new round of translation initiation. (B) Diagram showing the cyclical VLCFA elongase pathway and the genes required for the catalytic steps. Also depicted is the pathway utilizing VLCFAs by the ceramide synthase complex LAC1/LAG1 and LIP1 to make ceramide. Ceramide is later modified to generate various sphingolipids.

LIP1 is a component of the ceramide synthase complex required for the formation of ceramide from a VLCFA and a sphingoid base (34). Each sphingolipid contains one 24- to 26-carbon-long VLCFA in addition to the long-chain base and head group (35).

The ER in budding yeast is composed of the classical membrane network connected to the nuclear envelope as well as a network of tubules known as the cortical ER. The cortical ER extends throughout the cell and encases the inner face of the entire plasma membrane (36). In microscopy, the cortical ER can often be mistaken as the plasma membrane itself (36). Although the bulk of yeast's cortical ER lies under the plasma membrane, in most metazoan cells, including mammalian cells, the ER is continuous with the nuclear envelope and forms a network of tubules throughout the cytoplasm that closely align with microtubules (37).

Using protein affinity purifications coupled with mass spectrometry and yeast two-hybrid analysis, we provide direct evidence for an interaction between the *S. cerevisiae* eIF2B complex and YBR159W. We found that in wild-type (WT) cells eIF2B colocalizes with lipid membranes and that this membrane colocalization is not altered in a *ybr159w*Δ strain. Our experiments show that a *ybr159w*Δ mutation causes eIF2B to appear as numerous foci. Although *ybr159w*Δ null cells have a lower rate of translation, the appearance of numerous eIF2B foci does not appear to correlate with the cell's translation rate. Other VLCFA mutant strains showing multiple eIF2B foci have WT translation rates. Overall, we demonstrate here a novel interaction between the essential yeast translation initiation factor and the fatty acid synthesis enzyme YBR159W.

## MATERIALS AND METHODS

**Strains and media.** All yeast medium, growth, and genetic manipulation experiments were performed according to standard techniques (38). To create the *ybr159w*Δ strain AL401, the kanamycin resistance cassette from plasmid pFa6a-kanmx6 was first amplified with primers CGTACGCTGC AGGTCGAC and ATCGATGAATTCGAGCTCG. Using the PCR double-fusion approach (38), the primers CGGATTTGGAAGTCCTTTA TAG, GTCGACCTGCAGCGTACGCATTTCTTAAGCTGCACCG, CGA GCTCGAATTCATCGATTAGAATTATCGTTCTCG, and GGACTTGG TCCTTCCACC were used to expand the YBR159W genomic regions flanking the kanmx6 cassette. The YBR159W disruption cassette was transformed into strain BY4741, and transformants were selected on yeast extract-peptone-dextrose (YPD) plus 300 mM G418 plates and screened using Western blotting and anti-YBR159W polyclonal antibodies. Candidate BY4741 *ybr159w*Δ strains were crossed with the HIS$^+$ strain H1511 and sporulated to create the *ybr159w*Δ null strain AL401. An isogenic WT HIS$^+$ control strain AL400 was selected from the same sporulation. The *lip1*Δ strain RH5994 was kindly provided by the Howard Riezman laboratory (34). The *gcn3*Δ, *fen1*Δ, and *sur4*Δ deletion strains were obtained from the *MAT***a** yeast deletion collection (39). The *fen1*Δ and *sur4*Δ deletion strains from the *MAT***a** yeast deletion collection were mated with the HIS$^+$ strain H1511 and sporulated to create the *fen1*Δ and *sur4*Δ strains, AL413 and AL414, respectively. The tandem affinity purification (TAP)-tagged strains were obtained from the yeast TAP-tagged library (40). The GFP-tagged strains were obtained from a GFP-tagged yeast library (41). To make the *ybr159w*Δ, GCD7-GFP strain, we mated the *ybr159w*Δ strain AL401 with the GCD7-GFP strain AL429 from the GFP-tagged yeast library and sporulated the diploids to obtain the *ybr159w*Δ, GCD7-GFP strain AL403. The yeast two-hybrid activation-domain strains derived from parent strain PJ69-4a, the binding-domain parent strain PJ69-4α, and yeast two-hybrid plasmids were obtained from the Yeast Resource Center (University of Washington) (42). Using the protocol previously described by the Yeast Resource Center (42), the AL408 (YBR159W-BD), AL409 (GCD1-BD), AL410 (GCD2-BD), AL411 (GCD6-BD), and AL412 (GCD7-BD) strains expressing yeast two-hybrid binding-domain tagged alleles were generated from parent strain PJ69-4α. Briefly, initial forward and reverse primers were used to PCR the target gene from yeast genomic DNA. The PCR product and the common forward and reverse two-hybrid primers were used for a second round of PCR to extend the flanking sequences. The second PCR product and the PvuII- and NcoI-linearized pOBD2 plasmid were cotransformed into yeast strain PJ69-4α and plated on synthetic complete medium lacking Trp (SC-Trp) to select for recombinants fusing the target gene to the GAL4-binding domain. Tables 1, 2, and 3 list all of the strains, plasmids, and primers used in the present study.

**Plasmids.** The plasmid pOBD2 used in generating yeast two-hybrid binding-domain strains has been previously described (43). To create a plasmid expressing an endogenous level of YBR159W, we used PCR to amplify the YBR159W gene along with 600 bp of the genomic region upstream of the gene's start codon and the YBR159's stop codon using the primers CACCATGGTTTTTGTGACTTTACCTATAAATAGT ACACAAC and CTATTCCTTTTTTAACCTGTCTTGCGGCTTTTTTTA

TABLE 1 Strains used in this study

| Strain | Genotype | Source or reference |
|---|---|---|
| AL400 | *MATa ura3 leu2 HIS*$^+$ | This study |
| AL401 | *MATa ura3 leu2 HIS*$^+$ *YBR159W::Kan*$^R$ | This study |
| AL402 | *MATa ura3 leu2 HIS*$^+$ *YBR159W::Kan*$^R$ [*YCp-YBR159W*] | This study |
| AL403 | *MATa ura3 leu2 HIS*$^+$ *YBR159W::Kan*$^R$ *GCD7-GFP* | This study |
| AL404 | *MATa ura3 leu2 HIS*$^+$ *YBR159W::Kan*$^R$ *GCD7-GFP* [*YCp-YBR159W*] | This study |
| AL405 | *MATa leu2 ura3 met15 GCD1-GFP* [*YCp-YBR159W-dsRed*] | This study |
| AL406 | *MATa leu2 ura3 met15 GCD6-GFP* [*YCp-YBR159W-dsRed*] | This study |
| AL407 | *MATa leu2 ura3 met15 GCD7-GFP* [*YCp-YBR159W-dsRed*] | This study |
| AL408 | *MATα trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*YBR159W-GAL4DBD*] | This study |
| AL409 | *MATα trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD1-GAL4DBD*] | This study |
| AL410 | *MATα trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD2-GAL4DBD*] | This study |
| AL411 | *MATα trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD6-GAL4DBD*] | This study |
| AL412 | *MATα trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD7-GAL4DBD*] | This study |
| AL413 | *MATa leu2 ura3 met15 HIS*$^+$ *FEN1::Kan*$^R$ | This study |
| AL414 | *MATa leu2 ura3 met15 HIS*$^+$ *SUR4::Kan*$^R$ | This study |
| AL415 | *MATa ura3 leu2 HIS*$^+$ [p180] | This study |
| AL416 | *MATa ura3 leu2 HIS*$^+$ *YBR159W::Kan*$^R$ [p180] | This study |
| AL417 | *MATa ura3 leu2 HIS*$^+$ *YBR159W::Kan*$^R$ [*YCp-YBR159W*] [p180] | This study |
| AL418 | *MATa leu2 ura3 met15 FEN1::Kan*$^R$ [p180] | This study |
| AL419 | *MATa leu2 ura3 met15 SUR4::Kan*$^R$ [p180] | This study |
| AL420 | *MATα ura3-52 gcd1-101* [p180] | This study |
| AL421 | *MATα ura3-52 trp1-63 leu2-3 leu2-112 GAL2*$^+$ *gcn2Δ* [p180] | This study |
| AL422 | *MATa leu2 ura3 met15 FEN1-GFP* [*YCp-YBR159W-dsRed*] | This study |
| AL423 | *MATa leu2 ura3 met15 DPM1-GFP* [*YCp-YBR159W-dsRed*] | This study |
| AL424 | *MATa leu2 ura3 met15 his3 GCN3::Kan*$^R$ | Deletion library (39) |
| AL425 | *MATa leu2 ura3 met15 YBR159W-GFP* | GFP library (41) |
| AL426 | *MATa leu2 ura3 met15 FEN1-GFP* | GFP library (41) |
| AL427 | *MATa leu2 ura3 met15 GCD1-GFP* | GFP library (41) |
| AL428 | *MATa leu2 ura3 met15 GCD6-GFP* | GFP library (41) |
| AL429 | *MATa leu2 ura3 met15 GCD7-GFP* | GFP library (41) |
| AL430 | *MATa leu2 ura3 met15 GCD2-TAP* | TAP library (40) |
| AL431 | *MATa leu2 ura3 met15 GCD7-TAP* | TAP library (40) |
| AL432 | *MATa leu2 ura3 met15 YBR159W-TAP* | TAP library (40) |
| AL433 | *MATa leu2 ura3 met15 FEN1-TAP* | TAP library (40) |
| AL434 | *MATa leu2 ura3 met15 SUR4-TAP* | TAP library (40) |
| AL435 | *MATa leu2 ura3 met15 TSC13-TAP* | TAP library (40) |
| AL436 | *MATa leu2 ura3 met15 his3 FEN1::Kan*$^R$ | Deletion library (39) |
| AL436 | *MATa leu2 ura3 met15 DPM1-GFP* | GFP library (41) |
| AL437 | *MATa leu2 ura3 met15 his3 SUR4::Kan*$^R$ | Deletion library (39) |
| F98 | *MATα ura3-52 gcd1-101* | A. Hinnebusch |
| H1511 | *MATα ura3-52 trp1-63 leu2-3 leu2-112 GAL2*$^+$ | A. Hinnebusch |
| H2557 | *MATα ura3-52 trp1-63 leu2-3 leu2-112 GAL2*$^+$ *gcn2Δ* | A. Hinnebusch |
| pAD(GCD7) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD7-AD*] | Yeast Resource Center (42) |
| PJ69-4a | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* | Yeast Resource Center (42) |
| PJ69-4α | *MATα trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* | Yeast Resource Center (42) |
| pOAD(GCD1) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD1-AD*] | Yeast Resource Center (42) |
| pOAD(GCD2) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD2-AD*] | Yeast Resource Center (42) |
| pOAD(GCD6) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCD6-AD*] | Yeast Resource Center (42) |
| pOAD(GCN3) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*GCN3-AD*] | Yeast Resource Center (42) |
| pOAD(SUI2) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*SUI2-AD*] | Yeast Resource Center (42) |
| pOAD(TDH1) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*TDH1-AD*] | Yeast Resource Center (42) |
| pOAD(YBR159W) | *MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ* [*YBR159W-AD*] | Yeast Resource Center (42) |
| RH5994 | *MATα leu2 ura3 trp1 bar1 LIP1::HIS3* | H. Riezman |

AGGC. The PCR product was cloned into the pENTR entry vector using Directional TOPO Cloning (Invitrogen) to create pENTR-YBR159W 5′ untranslated region (UTR)-YBR159W. The YBR159W cassette was transferred to the pAG415GAL-ccdB yeast destination vector using LR Clonase recombination (Invitrogen) (44). To eliminate possible promoter interference, the vector's endogenous GAL promoter was deleted using the restriction enzymes SacI and SpeI and replaced with the primer insert GGGAGCTCCATACTGATTAGTACACTAGTGG and CCACTAGTGT ACTAATCAGTATGGAGCTCCC to create the YBR159W expression plasmid YCp-YBR159W. To create a plasmid expressing RFP-tagged YBR159W, the YBR159W open reading frame (ORF) without the stop codon was amplified by PCR and cloned into the pENTR vector creating pENTR-YBR159W. The YBR159W ORF insert was transferred by recombinational cloning into the pAG415GPD-ccdB-dsRed vector (Addgene) to create the final expression plasmid YCp-YBR159W-dsRed. All plasmids used in the present study are listed in Table 2.

**Antibodies.** The anti-YBR159W polyclonal antibodies were generated by inoculation of a rabbit with the synthetic peptide CETVKAENKKS GTRG (Covance). The peptide was covalently bound to cyanogen bromide beads (Sigma-Aldrich) to affinity purify anti-YBR159W from rabbit whole blood. Polyclonal antibodies to yeast SUI2 were kindly provided by Tom Dever. Polyclonal antibodies to yeast GCD6 and GCD1 were kindly provided by Alan Hinnebusch. The mouse anti-DPM1 was obtained from Molecular Probes. Antibodies to yeast TDH1, TDH2, and TDH3 were obtained from Millipore. The antibody to yeast RPL32 was kindly provided by Jonathan Warner.

TABLE 2 Plasmids used in this study

| Plasmid | Backbone | Notes | Source (reference) |
|---|---|---|---|
| pOBD2 | pOBD2 | $amp^R$ TRP1 CEN4 ORI GAL4-DBD | Yeast Resource Center (42) |
| YCp-YBR159W | pAG415GAL-ccdB | $amp^R$ LEU2 CEN ORI YBR159W 5′ UTR-YBR159W | This study |
| YCp-YBR159W-dsRed | pAG415GPD-ccdB-dsRed | $amp^R$ LEU2 CEN ORI $P_{GPD}$-YBR159W-dsRed | This study |
| p180 | YCp50 | $amp^R$ URA3 CEN ORI GCN4 5′ UTR-LacZ | A. Hinnebusch |
| pFa6a-kanmx6 | pFa6a-kanmx6 | $amp^R$ KanR2 | Addgene |
| pENTR | pENTR | $Kan^R$ | Invitrogen |
| pENTR-YBR159W 5′ UTR-YBR159W | pENTR | $Kan^R$ YBR159W 5′ UTR-YBR159W | This study |
| pENTR-YBR159W | pENTR | $Kan^R$ YBR159W | This study |
| pAG415GAL-ccdB | pAG415GAL-ccdB | $amp^R$ LEU2 CEN ORI ccdB | Addgene |
| pAG415GPD-ccdB-dsRed | pAG415GPD-ccdB-dsRed | $amp^R$ LEU2 CEN ORI ccdB-dsRed | Addgene |

**Mass spectrometry and proteomics.** For yeast TAP experiments, TAP-tagged protein complexes were purified as previously described (45, 46). For each TAP strain, a 2-liter culture was grown to an optical density at 660 nm ($OD_{660}$) of 1 to 2 in YPD. The purified TAP complexes were reduced with 1/10 volume of 50 mM dithiothreitol (DTT) at 65°C for 5 min, and cysteines were alkylated with 1/10 volume of 100 mM iodoacetamide at 30°C for 30 min. The proteins were digested overnight at 37°C with modified sequencing grade trypsin at an ~25:1 substrate/enzyme ratio (Promega, Madison, WI). Proteins were identified using multidimensional protein identification technology (MudPIT) and an LTQ linear ion trap mass spectrometer (Thermo Fisher) (47, 48). A fritless, microcapillary (100-μm-inner-diameter) column was packed sequentially with 12 cm of 5-μm $C_{18}$ reversed-phase packing material (Synergi 4μ Hydro RP80a; Phenomenex) and 3 cm of 5-μm strong cation exchange packing material (Partisphere SCX; Whatman). The entire trypsin-digested samples were loaded onto the biphasic column equilibrated in 0.1% formic acid–2% acetonitrile, which was then placed in-line with an LTQ linear ion trap mass spectrometer. An automated six-cycle multidimensional chromatographic separation was performed using buffer A (0.1% formic acid, 5% acetonitrile), buffer B (0.1% formic acid, 80% acetonitrile), and buffer C (0.1% formic acid, 5% acetonitrile, 500 mM ammonium acetate) at a flow rate of 300 nl/min. Cycles 1 to 6 consisted of 3 min of buffer A, 2 min of 0 to 100% buffer C, and 5 min of buffer A, followed by a 60-min linear gradient to 60% buffer B. In cycles 1 to 6, the percentage of buffer C was increased incrementally from 0, 15, 30, 50, 70, and 100% in each cycle. During the linear gradient, the eluting peptides were analyzed by one full mass spectrometry (MS) scan (300 to 2,000 *m/z*), followed by five tandem MS (MS/MS) scans on the five most abundant ions detected in the full MS scan while operating under dynamic exclusion.

For proteomic analysis of membrane float experiment's membrane and cytoplasmic fractions, a modified MudPIT protocol was utilized. Purified yeast protein and subcellular complexes were processed and analyzed essentially as described above except a 12-step MudPIT experiments was used with the salt pulses of 0, 25, 50, 75, 100, 150, 200, 250, 300, 500, and 750 mM and 1 M ammonium acetate. Eluting peptides were analyzed using an LTQ-OrbitrapXL mass spectrometer with preview mode and monoisotopic precursor selection enabled. The top 10 precursor ions based on intensity were fragmented using CID in the ion trap using 35% normalized collision energy. Dynamic exclusion was enabled for 180 s with repeat count of 1 at 30-s duration, list size of 500, mass tolerance of 10 ppm. MS data were analyzed as previously described (49).

**GFP affinity purification.** Two liters of the GCD7-GFP *ybr159wΔ* strain AL403, the GCD7-GFP strain AL429, and the untagged *ybr159wΔ* strain AL401 were grown to an $OD_{660}$ of 1 to 2 in YPD. Yeast cells were harvested by centrifuging at 1,500 × *g* for 5 min and resuspended in 10 ml

TABLE 3 Primers used in this study

| Primer | Orientation[a] | Sequence (5′–3′) |
|---|---|---|
| Genomic YBR159W deletion primers | | |
| YBR159W deletion primer | F | CGTACGCTGCAGGTCGAC |
| | R | ATCGATGAATTCGAGCTCG |
| 5′ Homology extension primer | F | CGGATTTGGAAGTCCTTTATAG |
| | R | GTCGACCTGCAGCGTACGCATTTCTTAAGCTGCACCG |
| 3′ Homology extension primer | F | CGAGCTCGAATTCATCGATTAGAATTATCGTTCTCG |
| | R | GGACTTGGTCCTTCCACC |
| Yeast two-hybrid primers | | |
| Two-hybrid common primer | F | CTATCTATTCGATGATGAAGATACCCCACCAAACCCAAAAAAAGAGATCGAATTCCAGCTGACCACCATG |
| | R | GTACCGTTAAGGGCCCCTAGGCAGCTGGACGTCTCTAGATACTTAGCATCTATGACTTTTTGGGGCGTTC |
| Two-hybrid YBR159W | F | AATTCCAGCTGACCACCATGACTTTTATGCAACAGCTTCAAGAGGCTGG |
| | R | GATCCCCGGGAATTGCCATGCTATTCCTTTTTAACCTGTCTTGCGGCTTTTTTTAAGG |
| Two-hybrid GCD1 | F | AATTCCAGCTGACCACCATGTCAATTCAGGCTTTTGTCTTTTGCGGTAAAGG |
| | R | GATCCCCGGGAATTGCCATGTTAACGCTCAAATAATCCGTCATCTTCGTACTCGTAC |
| Two-hybrid GCD2 | F | AATTCCAGCTGACCACCATGAGCGAATCGGAAGCCAAATCTAGGTCG |
| | R | GATCCCCGGGAATTGCCATGTTATGCGGAACCTTTGTACTCTCTTAAAATAACAGGGAC |
| Two-hybrid GCD6 | F | AATTCCAGCTGACCACCATGGCTGGAAAAAAGGGACAAAAGAAAAGTGGACTAG |
| | R | GATCCCCGGGAATTGCCATGTTATTCCTCTTCTGAGGAAGATTCTTCGTCAGCATTC |
| Two-hybrid GCD7 | F | AATTCCAGCTGACCACCATGTCCTCTCAAGCATTCACTTCAGTACATCCG |
| | R | GATCCCCGGGAATTGCCATGTCACGCCTTATTTTTATCCAAATGCACATCAATTTGC |
| YBR159W ORF + 600-bp upstream primers | | |
| 600-bp upstream YBR159W primer | F | CACCATGGTTTTTGTGACTTTACCTATAAATAGTACACAAC |
| YBR159W ORF primer | R | CTATTCCTTTTTAACCTGTCTTGCGGCTTTTTTTAAGGC |
| pAG415GAL-ccdB promoter remover 1 | | GGGAGCTCCATACTGATTAGTACACTAGTGG |
| pAG415GAL-ccdB promoter remover 2 | | CCACTAGTGTACTAATCAGTATGGAGCTCCC |

[a] F, forward; R, reverse.

of ice-cold NP-40 lysis buffer (6 mM $Na_2HPO_4$, 4 mM $NaH_2PO_4$, 1% NP-40, 150 mM NaCl, 2 mM EDTA, 50 mM NaF, 4 μg of leupeptin/ml, 0.1 mM $Na_3VO_4$). Cells were lysed for 10 min with glass beads in NP-40 lysis buffer. The lysates were centrifuged at 500 × *g* for 5 min. The cleared supernatant was brought up to 25 ml with ice-cold lysis buffer. A 500-μl bed volume of protein A/G-agarose beads (Thermo Scientific) and 50 μg of anti-GFP antibody (ThermoFisher) were added simultaneously and allowed to incubate for 1 h at room temperature. The beads were centrifuged at 300 × *g* for 5 min, transferred to a Poly Prep chromatography column (Bio-Rad), and washed at 4°C with 50 column volumes of wash buffer (10 mM Tris [pH 8.0], 150 mM NaCl, 0.1% NP-40). Protein digestion was carried out directly on the agarose beads. The beads were suspended in 1 ml of digestion buffer (10 mM Tris [pH 8.0], 150 mM NaCl, 0.5 mM EDTA, 1 mM DTT) and transferred to a 1.5-ml microcentrifuge tube. The resuspended beads were trypsin digested as described for yeast TAP complexes. After digestion the beads were centrifuged at 13,000 × *g* for 1 min, and the supernatant was transferred to a fresh microcentrifuge tube. MudPIT was performed identical as described for the TAP purifications. MS data was analyzed as previously described using $C_n$ scoring filters of 1.5 (+1), 3.5 (+2), and 3.5 (+3) (49).

**Fatty acid profiling.** The protocol for extracting lipids from yeast cells was adapted from Ejsing et al. (50). Each yeast strain was grown to an $OD_{660}$ of 1.0 to 1.5 in YPD medium at 30°C. A 50-mg portion of wet weight yeast cells was incubated in 200 μl of phosphate-buffered saline (PBS) with 100 μg of lyticase (Sigma)/ml for 1 h at 37°C. Next, 990 μl of chloroform-methanol (17:1 [vol/vol]) was added, followed by incubation for 2 h at 37°C. The lower organic layer was collected and vacuum evaporated. Next, 990 μl of chloroform-methanol (2:1 [vol/vol]) was added to the upper aqueous layer and incubated for 2 h at 37°C. The lower layer was collected and pooled with the evaporated fraction taken from the first extraction and vacuum evaporated. The sample was solubilized with 100 μl of chloroform-methanol (1:2 [vol/vol]) and mixed 1:1 with 0.4 mM methylamine in methanol. Samples were directly injected into an ESI-LTQ OrbitrapXL at 2 μl/min and precursor ions were scanned using the Orbitrap analyzer at a resolution of 30,000 in negative ion mode. Using published inositolphosphoceramide (IPC) precursor *m/z* values, precursor ion peaks were identified using a mass tolerance of 10 ppm (51, 52). Using IPC structure data at the LIPID MAPS Lipidomics Gateway (52), the following theoretical precursor $[M-H]^-$ ion *m/z* values were used to identify the IPC ions in the high-resolution scan (sphingolipid species are identified by lipid class followed by numbers indicating carbons in FA moiety:double bonds in FA moiety;hydroxyl groups in FA moiety): IPC 46:0;4, 980.717); IPC 44:0;4, 952.686; IPC 42:0;4, 924.655; IPC 40:0;4, 896.623; IPC 38:0;4, 868.592. To validate the identity of these IPC ions, the IPC precursor ions were fragmented by CID in the linear ion trap. The observed *m/z* values of the MS/MS fragment ions for each IPC precursor was compared to predicted $[\text{ceramide phosphate-}H_2O]^-$ and $[\text{ceramide phosphate}]^-$ *m/z* values at a mass tolerance of 0.1 Da. The following theoretical *m/z* $[\text{ceramide phosphate-}H_2O]^-$ and $[\text{ceramide phosphate}]^-$ fragment ion values were used to validate the IPC lipids: IPC 38:0;4, 688.53, 706.54; IPC 40:0;4, 716.56, 734.57; IPC 42:0;4, 744.59, 762.60; IPC 46:0;4, 800.65, 818.66. In addition, to validate the identification of IPC 44:0;4, the fragmentation spectrum of the precursor with *m/z* 952.68 at a mass tolerance of 0.1 Da was compared to the previously published values for the fragment ions $[\text{ceramide phosphate-}H_2O]^-$, *m/z* 772.62, and $[\text{ceramide phosphate}]^-$, *m/z* 790.63 (51). To compare the observed abundance for each IPC species between strains, the precursor ion signal intensity for each identified IPC species was normalized to the signal intensity of the *m/z* 835.53 base peak corresponding to the phosphatidylinositol (PI) species PI 16:1-18:0 and PI 16:0-18:1, where the numbers indicate carbons in first FA moiety:double bonds in first FA moiety-carbons in second FA moiety:double bonds in second FA moiety.

**Growth rate analysis.** Yeast strains were grown overnight at 30°C in YPD. Relative cell number was measured as the $OD_{660}$ by using a Beckman DU 530 spectrophotometer. Cells were diluted in 50 ml of fresh YPD to ~0.05 $OD_{660}$ unit/ml. Individual strains were grown at 30°C, and an $OD_{660}$ measurement was obtained every 2 h. The formula for used for converting $OD_{660}$ readings to cell numbers was $y = 1.1564x^3 - 0.6815x^2 + 1.3996x$, with *y* as cell number/ml and *x* as the $OD_{660}$ value (38). Cell doubling time was determined by plotting the growth curve for each strain and measuring the maximum rate of cell growth during logarithmic growth.

**Yeast two-hybrid assay.** Mating type A strains containing AD-tagged alleles and mating type α strains containing BD-tagged alleles have been previously described (53). The A and α strains were allowed to mate in liquid YPD at 30°C overnight. The relative cell number was determined by measuring the $OD_{660}$, and 4 μl of a solution containing $10^7$ cells/ml was plated onto SC-Leu-Trp-His–1.5 mM 3-aminotriazole (3-AT) agar plates. The plates were scanned after 48 h.

**Membrane flotation.** Membrane flotation of yeast extracts was performed as previously described (54). A 50-ml culture of each yeast strain was grown to an $OD_{660}$ of 1.0 to 1.5 in YPD medium at 30°C. The cells were lysed with glass beads in ice-cold breaking buffer (30 mM Tris [pH 7.0], 1 mM EDTA). The lysate was cleared by centrifugation at 500 × *g* for 3 min. Lysate corresponding to 10 $OD_{660}$ units of cells in 222 μl was mixed with 1,778 μl of ice-cold 90% sucrose (wt/vol)–10 mM Tris (pH 7.0) solution. Then, 2 ml of lysate-sucrose solution was transferred to the bottom of a 14- by 89-mm ultracentrifuge tube (Beckman, catalog no. 344059) and layered with 6 ml of 65% sucrose–10 mM Tris (pH 7.0) and then 3 ml of 10% sucrose–10 mM Tris (pH 7.0). The tubes were centrifuged in a Beckman SW-41 rotor at 24,000 rpm for 18 h. Individual 1.5-ml fractions were collected from the top of the gradient, and the proteins were trichloroacetic acid (TCA) precipitated. Ten percent of each fraction was used for SDS-PAGE and Western blotting.

**Membrane flotation fractions affinity purifications.** For each TAP strain, a 1-liter culture was grown to an $OD_{660}$ of ~1 in YPD, and the cells were split into six fractions. Each cell fraction was separated using the membrane flotation gradients as described above. The 10 to 65% sucrose interface layer and a 80% sucrose layer from each gradient were collected and pooled. TAP purification was performed as previously described up to TEV protease cleavage (45, 46).

**GCN4-LacZ induction.** The yeast reporter plasmid p180 containing the GCN4 5′ UTR coupled to a *lacZ* reporter has been previously described (17). Yeast strains transformed with p180 were grown overnight at 30°C in SC-uracil (SC-ura). Cultures were diluted 1:10 and allowed to continue growing for 2 h in SC-ura-His. Cells were spun down and split into two tubes containing 10 ml of SC-ura-His medium. A 1 M 3-AT solution was added to the starvation tube to a final concentration of 10 mM. The cells continued to grow for 4 h at 30°C. β-Galactosidase assays were performed as previously described (55). The cells were centrifuged at 1,500 × *g* for 5 min and lysed with glass beads in 1 ml of ice-cold breaking buffer (100 mM Tris [pH 8.0], 1 mM DTT, 20% glycerol). A 20-μl portion of whole-cell extract was added to 900 μl of Z buffer (16.1 g of $Na_2HPO_4 \cdot 7H_2O$/liter, 5.5 g of $NaH_2PO_4 \cdot H_2O$/liter, 0.75 g of KCl/liter, 0.246 g of $MgSO_4 \cdot 7H_2O$/liter, 2.7 ml of β-mercaptoethanol/liter [pH 7.0]), followed by incubation at 28°C for 5 min. The reaction was initiated by adding 200 μl of 4 mg of ONPG (*o*-nitrophenyl-β-D-galactopyranoside)/ml in Z buffer, followed by incubation at 28°C. After the reaction turned a pale yellow color, 0.5 ml of 1 M $Na_2CO_3$ was added. LacZ expression was determined by measuring the absorbance at 420 nm using a Beckman DU 530 spectrophotometer. The protein concentration of the extracts was determined by using the Bio-Rad DC protein assay. LacZ specific activity was determined according to the following formula: $[(OD_{420} \times 1.7)/(0.0045 \times \text{protein concentration (mg/ml)} \times \text{extract volume (ml)} \times \text{time (min)})]$ (38). Values were normalized to the WT.

**[$^{35}$S]methionine incorporation.** Overnight cultures of yeast grown in YPD were diluted 1:10 in 10 ml of SC-Met and grown for 3 h at 30°C. The $OD_{660}$ of the culture was measured to determine the cell numbers. For labeling, [$^{35}$S]methionine (MP Biomedicals) was added to 5 ml of the cell culture to a final concentration of 10 μCi/ml. Samples were incubated

with shaking for 30 min at 30°C. Labeling was stopped by the addition of 1/10 volume 100% TCA and heating to 100°C for 30 min. TCA precipitates were collected on GFC filters (Whatman) and then washed sequentially with 5 ml each of 10% TCA and 95% ethanol. Filters were then placed in 5 ml of EcoLume scintillation fluid (MP Biomedicals), and [$^{35}$S]methionine incorporation was measured using a Beckman LS 6500 scintillation counter. Values were reported as the counts per minute/$OD_{660}$ unit.

**Microscopy.** Epifluorescence microscopy was performed using live yeast cells grown in SC medium to an $OD_{660}$ of 1.0 to 1.5 at 30°C. Cells were mounted on slides and visualized using a Zeiss Axiophot bright-field microscope with a ×63/1.40 Plan-Apochromat oil differential interference contrast (DIC) lens. Images were analyzed with MetaMorph imaging software (Molecular Devices). Live yeast cells imaged using confocal microscopy were grown in SC medium to an $OD_{660}$ of 1.0 to 1.5 at 30°C. Cells were visualized with a Zeiss LSM 510 META inverted confocal microscope using a ×63/1.40 Plan-Apochromat oil immersion lens. Microscopic images used for quantitative analysis were analyzed using ImageJ imaging software (56). To quantify the percentage of 2B bodies that colocalized with the ER, a 2B body was judged to be colocalized with the ER only if the 2B body signal overlapped with an area of YBR159W at least half as bright as the brightest YBR159W signal seen in the cell. Cells were pooled into groups of ~25 cells to calculate a standard deviation for the percentage of 2B bodies colocalized with the ER. The bright regions of the ER were subtracted from the total area of the cell minus the nuclear area to determine the fraction of the cell taken up by the ER. The compound 3,3′-dihexyloxacarbocyanine iodide [$DiOC_6(3)$] was used to stain and image the membranes of the WT strain AL400 and the *ybr159wΔ* strain AL401 as previously described (57). Yeast cells were incubated in media containing 2.5 μg of $DiOC_6(3)$/ml for 10 min before imaging.

**Polysome profiling.** Polysome analysis was performed as previously described (58). Yeast strains were grown in YPD to an $OD_{660}$ of ~1. Cells were lysed with glass beads in ice-cold breaking buffer (10 mM Tris [pH 7.0], 100 mM NaCl, 30 mM $MgCl_2$, 50 μg of cycloheximide/ml, 200 μg of heparin/ml). The crude lysate was cleared by centrifugation at 500 × *g* for 3 min, and 20 $OD_{660}$ units of cells were loaded on top of a 7 to 47% continuous sucrose gradient (wt/vol) cast in 50 mM Tris (pH 7.0), 50 mM $NH_4Cl$, 12 mM $MgCl_2$, and 50 μg of cycloheximide/ml in a 14-by-89-mm ultracentrifuge tube (Beckman). Gradients were centrifuged in a Beckman SW-41 rotor at 14,000 rpm for 18 h at 4°C. An absorbance profile at 254 nm was collected from the gradients as previously described (16). Fractions (1 ml) were used for Western blotting. Monosome and polysome peak areas were determined by using ImageJ software (56). A moving baseline for each profile was established by connecting the minima between each peak, and the area under each peak above this line was calculated. The polysome peak areas were summed and compared to the monosome peak area.

**Subcellular fractionation.** WT yeast strain AL400 was grown to an $OD_{660}$ of 1.0 to 1.5 in YPD medium at 30°C. To isolate subcellular fractions, 45 $OD_{660}$ units of cells were split into three samples: control, puromycin treatment, and EDTA treatment. The control sample was lysed using glass beads in 750 μl of ice-cold control buffer (10 mM Tris [pH 8.0], 100 mM NaCl, 1 mM EDTA, 10 mM KCl, 30 mM $MgCl_2$). The puromycin and EDTA treatment samples were lysed using glass beads in 750 μl of ice-cold ribosome dissociation buffer (10 mM Tris [pH 8.0], 100 mM NaCl, 1 mM EDTA). The control sample was diluted in 750 μl of control buffer. The puromycin treated sample was diluted with 750 μl of ribosome dissociation buffer containing 2 mM puromycin to a final concentration of 1 mM. The EDTA-treated sample was diluted in ribosome dissociation buffer (20 mM EDTA) to a final concentration of 10 mM EDTA. Lysates were gently mixed at room temperature for 30 min to facilitate dissociation of ribosomes from the ER. Lysates were centrifuged at 900 × *g* for 5 min, and the supernatant was centrifuged at 11,000 × *g* for 20 min. The soluble fraction was recovered from the supernatant. The pellets were resuspended in either control buffer (10 mM Tris [pH 8.0], 100 mM NaCl, 1 mM EDTA, 10 mM KCl, 30 mM $MgCl_2$), puromycin solution (1 mM puromycin, 10 mM Tris [pH 8.0], 100 mM NaCl, 1 mM EDTA), or EDTA solution (10 mM EDTA, 10 mM Tris [pH 8.0], 100 mM NaCl) and centrifuged at 11,000 × *g* for 20 min. The pellets were resuspended in 1.5 ml of resuspension buffer (1 mM puromycin, 10 mM Tris [pH 8.0], 100 mM NaCl, 1 mM EDTA, 1% SDS), and 15 μl of each fraction was used for Western blotting.

## RESULTS

In a tandem affinity purification proteomics screen of *S. cerevisiae* translation initiation factors, followed by liquid MS analysis, we discovered that all five subunits of eIF2B copurified with the VLCFA enzyme YBR159W (A. J. Link et al., unpublished data) (Fig. 2A). Subsequent liquid chromatography (LC)-MS/MS analysis of TAP-YBR159W affinity purification showed YBR159W copurified with all five subunits of the eIF2B complex and several members of the VLCFA synthesis pathway. In the present study, additional TAP experiments examined whether other members of the VLCFA synthesis pathway also interact with eIF2B subunits. With the exception of YBR159W, our data showed that other members of the VLCFA synthesis pathway did not interact with eIF2B (Fig. 2A). To rule out the possibility that the YBR159W-eIF2B interaction was due to an artifact of the TAP-tagged strains, we performed a GFP affinity purification using the GCD7-GFP strain AL429. LC-MS/MS analysis identified YBR159W copurifying with all five subunits of eIF2B (see Fig. 5E). Next, we utilized yeast two-hybrid analysis to identify interactions between eIF2B subunits and YBR159W. The activation-domain tagged strains pOAD(YBR159W), pOAD(GCD1), pOAD(GCD2), pOAD(GCN3), pOAD(GCD6), pOAD(GCD7), pOAD(SUI2), and pOAD(TDH1) were mated with binding-domain tagged strains AL408 (YBR159W), AL409 (GCD1), AL410 (GCD2), AL411 (GCD6) and AL412 (GCD7). The positive interactions between different subunits of the eIF2B complex validated the experiment's ability to detect previously described interactions (Fig. 2B). The two-hybrid analysis showed that YBR159W positively interacted with both the GCD6 and GCD7 subunits of eIF2B (Fig. 2B).

The GFP-tagged YBR159W strain AL425 showed the YBR159W protein localizes to the ER membrane using epifluorescence microscopy (Fig. 3A). DPM1 encodes the enzyme dolichol phosphate mannose synthase that adds a mannose moiety to dolichyl phosphate on the cytosolic side of the endoplasmic reticulum (59, 60). DPM1 is an ER membrane protein unrelated to VLCFA synthesis or utilization (60). Confocal microscopy using the FEN1-GFP YBR159W-RFP strain AL422 and the DPM1-GFP YBR159W-RFP strain AL423 confirmed that RFP-tagged YBR159W expressed from a low-copy-number plasmid colocalizes with the VLCFA protein FEN1 and ER protein DPM1 (Fig. 3B).

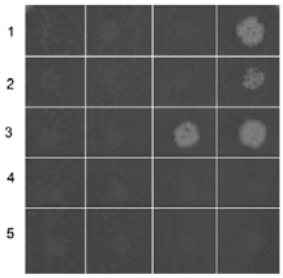
We constructed the *ybr159wΔ* yeast strain AL401 to examine the null phenotype. The mutant strain had a slow-growth phenotype (Fig. 3C) and was temperature sensitive at 37°C (data not shown). To show that the slow-growth phenotype was due to the deletion of *ybr159wΔ* and not a second site mutation in the strain, the *ybr159wΔ* null yeast strain was complemented in strain AL402 expressing YBR159W from the low-copy-number plasmid YCp-YBR159W (Fig. 3C). Our results agreed with previous studies using an unrelated *ybr159wΔ* null strain (29).

Previous work has shown that disruption of VLCFA utilization in yeast causes abnormal formation of lipid membranes (61). The compound 3,3′-dihexyloxacarbocyanine iodide [$DiOC_6(3)$] is a

A

| Gene | TAP-GCD2 | TAP-YBR159W | TAP-FEN1 | TAP-SUR4 | TAP-TSC13 |
|---|---|---|---|---|---|
| GCD1 | 10(28.72) | 10(23.49) | - | - | - |
| GCD2 | 7(20.58) | 10(21.47) | - | - | - |
| GCN3 | 7(47.54) | 4(17.97) | - | - | - |
| GCD6 | 13(30.34) | 16(30.58) | - | - | - |
| GCD7 | 8(34.91) | 6(21.47) | - | - | - |
| | | | | | |
| YBR159W | 6(23.92) | 10(27.59) | - | - | - |
| FEN1 | - | - | 2(6.63) | - | - |
| SUR4 | - | - | - | 2(5.20) | 1(5.49) |
| TSC13 | - | 1(2.89) | - | 1(3.22) | 2(5.79) |

B




| | A | B | C | D |
|---|---|---|---|---|
| 1 | YBR : GCD1 | YBR : GCD2 | YBR : GCD6 | YBR : GCD7 |
| 2 | GCD1 : YBR | GCD2 : YBR | GCN3 : YBR | GCD6 : YBR |
| 3 | GCD7 : YBR | GCD6 : GCD1 | GCD6 : GCD7 | GCD1 : GCD7 |
| 4 | GCD1 : GCD6 | GCD2 : GCD6 | GCN3 : GCD6 | GCD7 : GCD6 |
| 5 | SUI2 : GCD1 | SUI2 : YBR | TDH1 : GCD1 | TDH1 : YBR |

KEY: BAIT : PREY; YBR = YBR159W; Positive interaction

FIG 2 YBR159W's interaction with eIF2B is unique among VLCFA genes. (A) MS analysis of the affinity-purified TAP-GCD2, TAP-YBR159W, and other TAP-tagged VLCFA protein complexes. Listed are unique peptide identifications with the percent coverage of identified peptides in the protein in parentheses. A "–" indicates that no peptides were identified for the gene. (B) Yeast two-hybrid (Y2H) analysis of interactions between YBR159W and eIF2B subunits GCD6 and GCD7. Shown is both the assay plate used for scoring the Y2H interactions and a table of the interactions tested at each spot. Shading on the table corresponds to a positive interaction on the plate.

lipophilic dye used to label a variety of lipid membranes (57). We used $DiOC_6(3)$ to stain membranes of WT strain AL400, *ybr159w*Δ strain AL401, and the VLCFA mutant strains AL413 (*fen1*Δ), and AL414 (*sur4*Δ), and the ceramide synthase mutant strain RH5994 (*lip1*Δ). The *ybr159w*Δ, *fen1*Δ, *sur4*Δ, and *lip1*Δ null strains all displayed disrupted lipid membranes using epifluorescence imaging (Fig. 4). This supported previous work showing that VLCFAs are important for proper membrane formation (61).

To determine whether YBR159W has a role in translation, we examined whether the *ybr159w*Δ strain AL401 causes a defect in protein synthesis. We used [$^{35}$S]methionine incorporation to quantify the global translation rate. The [$^{35}$S]methionine incorporation experiments showed that the *ybr159w*Δ strain has a reduced translation rate (Fig. 5A). The ceramide synthase mutant *lip1*Δ strain RH5994 also showed a reduction in the rate of translation. The *lip1*Δ strain had a slow growth rate similar to that of the *ybr159w*Δ strain. However, the VLCFA mutant strains AL413 (*fen1*Δ) and AL414 (*sur4*Δ) showed no reduction in translation or growth rates (Fig. 5A).

Next, we performed polyribosome profiling to examine the distribution of 40S, 60S, 80S, and polyribosomes in the *ybr159w*Δ strain AL401. Compared to the WT strain, we observed the polysome profiles for the *ybr159w*Δ strain showed an increase in the 80S monosome peak and a decrease in polysome peaks (Fig. 5B). As expected, the complemented *ybr159w*Δ strain AL402 showed a polysome profile similar to that of the WT. To normalize and quantify the observed differences in the peak areas, the ratio of the 80S monosome to polysome peak areas was calculated. The monosome/polysome ratio significantly increased for the *ybr159w*Δ strain compared to the WT and complemented strains (Fig. 5D). Polysome profiles of the *lip1*Δ strain RH5994 showed defects similar to those of the *ybr159w*Δ strain (Fig. 5B and D). Polysome profiling of the *fen1*Δ strain AL413 and *sur4*Δ strain AL414 showed no noticeable differences from WT strain AL400 (Fig. 5B and D). These polysome distributions were consistent with the reduced global translation rates seen previously in the [$^{35}$S]methionine labeling experiments.

We next examined *ybr159w*Δ's effect on eIF2B's activity. We used a GCN4-lacZ expression assay to examine GCN4 expression during the starvation response (62). Strains AL400 (HIS$^+$ control strain), AL401 (*ybr159w*Δ), AL402 (*ybr159w*Δ + YCp-YBR159W), AL413 (*fen1*Δ), AL414 (*sur4*Δ), RH5994 (*lip1*Δ), H2557 (*gcn2*Δ), and F98 (*gcd1*) were transformed with the GCN4-lacZ reporter plasmid p180. Our results showed that the *ybr159w*Δ, *fen1*Δ, *sur4*Δ, and *lip1*Δ null strains did not affect the induction of GCN4 during amino acid starvation (Fig. 5C). This suggested that eIF2B's role in the regulation of GCN4 response is not affected by the *ybr159w*Δ null or other VLCFA pathway mutation.

We next tested whether the *ybr159w*Δ mutation affected the composition of the eIF2B complex. Using the *ybr159w*Δ GCD7-

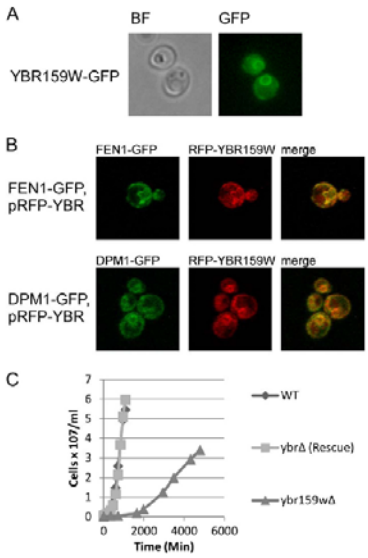



FIG 3 Cellular analysis of YBR159W. (A) Live cell epifluorescence imaging of endogenously tagged YBR159W-GFP indicates YBR159W localizes mainly to the ER membrane. (B) Live cell confocal microscopy showing the colocalization of YBR159W with the VLCFA pathway enzyme FEN1 and ER membrane protein DPM1. YBR159W is expressed on a low-copy-number plasmid and tagged with DsRed. FEN1 and DPM1 are endogenously expressed and tagged with GFP. (C) Deletion of YBR159W results in a very slow growth rate.

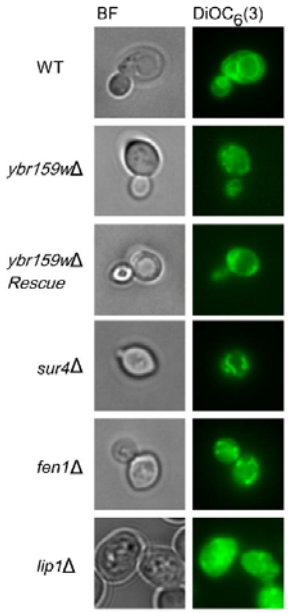
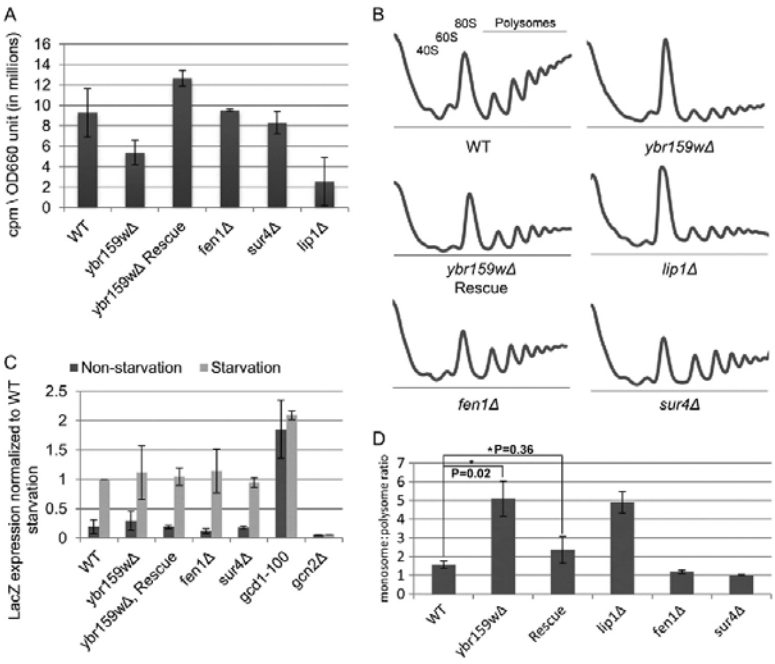



FIG 4 Null mutations of genes in the VLCFA pathway disrupt lipid membranes. The lipophilic dye $DiOC_6(3)$ was used to label membranes in live yeast cells. Dye was applied to cells in suspension 10 min before plating on a microscope slide and imaging. Included are the VLCFA and ceramide synthase *fen1*Δ, *sur4*Δ, and *lip1*Δ mutants as controls. A total of 100% of these mutants showed abnormal membranes ($n$ = 246) versus 1.1% for WT ($n$ = 89) and 9.7% for the rescue ($n$ = 93).

GFP tagged strain AL403, the untagged *ybr159w*Δ strain AL401, and the GCD7-GFP strain AL429, we performed GFP affinity purifications and LC-MS/MS analysis of the affinity-purified complexes. All five subunits of eIF2B were identified in the *ybr159w*Δ GCD7-GFP strain and the GCD7-GFP strain (Fig. 5E). No subunits of eIF2B were identified in the untagged *ybr159w*Δ control strain AL401. These results suggested that the composition of eIF2B is not dependent upon the presence of YBR159W.

Although the composition of eIF2B appeared to be independent of YBR159W, the consistently lower number of identified peptides for each eIF2B subunit from the MS data for the GCD7-GFP, *ybr159w*Δ null strain compared to the GCD7-GFP strain suggested that the cellular abundance of eIF2B was lower in the *ybr159w*Δ null background (Fig. 5E). To determine whether the cellular abundance of eIF2B is lower in a *ybr159w*Δ null strain, Western analysis was performed on the yeast strains used in the GCD7-GFP affinity purification of eIF2B complexes. Lack of signal for YBR159W in the *ybr159w*Δ strains confirmed the expected null genotype (Fig. 5F). In concordance with the MS results, the GCD7-GFP, *ybr159w*Δ strain had a lower abundance of eIF2B compared to the GCD7-GFP strain (Fig. 5E). To validate this observation in untagged strains, Western analysis was also performed using the WT strain AL400 and the untagged *ybr159w*Δ null strain AL401. The *ybr159w*Δ null strain again showed lower abundance of eIF2B compared to the WT strain (Fig. 5F)

We next tested whether eIF2B played a role in VLCFA synthesis. Previous studies had shown a *ybr159w*Δ null strain had an altered VLCFA lipid composition (30). Since four of the five subunits of eIF2B are essential, we used a *gcn3*Δ strain AL424 to test for VLCFA defects. The WT strain AL400, the *ybr159w*Δ strain AL401, the *ybr159w*Δ rescue strain AL402, the *sur4*Δ strain AL414, and the *lip1*Δ strain RH5994 were used as positive and negative controls. To profile the VLCFAs, lipids were extracted from yeast cells and directly infused into an ESI-LTQ-OrbitrapXL mass spectrometer while scanning at high resolution in negative-ion mode. Several IPCs, a class of VLCFA-containing sphingolipid, were identified using previously published *m/z* values at a 10-ppm mass accuracy (50, 52). We validated the identification of the IPC species using either previously observed fragmentation spectrum or expected *m/z* values for the IPC's $[\text{ceramide phosphate-}H_2O]^-$ and $[\text{ceramide phosphate}]^-$ fragment ions (Fig. 6) (51). The IPC 44:0;4 and IPC 46:0;4 sphingolipids contain full-length VLCFAs and are the most abundant yeast sphingolipid species (24). Compared to the WT strain, the *gcn3*Δ, *ybr159w*Δ, and other VLCFA and ceramide synthase mutant strains all showed a reduction in the IPC 44:0;4 and IPC 46:0;4 sphingolipids containing full-length VLCFAs (Fig. 7A). The IPC sphingolipid species IPC 38:0;4, IPC 40:0;4, and IPC 42:0;4 contain shorter-chain fatty acids and are typically only detected in VLCFA biosynthesis mutant strains (24). As previously observed, the *sur4*Δ strain had elevated shorter-chain fatty acid-containing IPC species 38:0;4, 40:0;4, and 42:0;4 (36). We observed that IPC 38:0;4 and IPC 42:0;4 were also elevated in the *ybr159w*Δ strain. The *gcn3*Δ strain showed no significant changes in the shorter-chain fatty acid sphinglipid's IPC




E

| Gene | *ybr159wΔ* | GCD7-GFP | GCD7-GFP, *ybr159wΔ* |
|---|---|---|---|
| YBR159W | - | 15(47.55) | - |
| GCD1 | - | 33(60.21) | 13(33.39) |
| GCD2 | - | 30(36.41) | 10(25.96) |
| GCN3 | - | 13(58.36) | 8(50.49) |
| GCD6 | - | 34(54.63) | 16(30.76) |
| GCD7 | - | 21(55.12) | 11(38.58) |

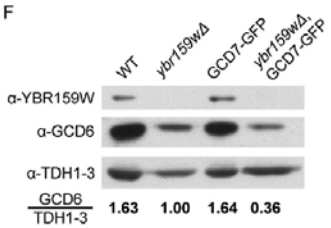



FIG 5 Translation assays on the *ybr159w*Δ strain. (A) Translation efficiency as measured by [$^{35}$S]methionine incorporation. Values are counts per minute per $OD_{660}$ unit of cells. The results shown are from at least three replicates. (B) Polysome profiling of the *ybr159w*Δ and other VLCFA null strains. At least three replicates were performed for each strain. Although the example *ybr159w*Δ plot does not show a 40S ribosome peak, all other replicates of the strain showed a 40S peak similar to the WT. (C) Assay for GCN4 pathway competence by GCN4-LacZ induction. The results are expressed as the LacZ expression per mg of protein per min normalized to the WT starvation condition. Starvation conditions were induced by 10 mM 3-AT in synthetic complete minus histidine media for 4 h. The *gcd1-100* strain has a constitutively derepressed GCN4 pathway and constant GCN4 protein translation, while the *gcn2*Δ strain is incapable of derepression of GCN4 and cannot produce significant amounts of GCN4 protein. (D) Ratio of monosome to polysome peak areas for the polysome profiles. *P* values were generated using the Student *t* test from at least three individual replicates. (E) GFP pulldown of eIF2B complexes in a *ybr159w*Δ background. After pulldown LC-MS/MS was performed to identify the proteins. An untagged *ybr159w*Δ strain and GCD7-GFP tagged strain were used as controls. Displayed are unique peptide hits and the percentage of coverage as described in Fig. 2A. (F) Western blot analysis of WT, *ybr159w*Δ, and GFP-GCD7 strains. The yeast strains in panel D and the WT strain AL400 were used. Equivalent amounts of whole-cell extracts were loaded on the SDS-PAGE gel. The Western signals for GCD6 and TDH1 to -3 were determined by densitometry. The ratio of the anti-GCD6 to the anti-TDH1-3 signal is shown for each strain. The anti-TDH1-3 antibody does not distinguish between the three GAPDH (glyceraldehyde-3-phosphate dehydrogenase) gene duplications in yeast.

38:0;4, 40:0;4, and 42:0;4 levels (Fig. 7B). The *lip1*Δ strain contained barely perceptible levels of any IPC, supporting its requirement for ceramide synthesis (34).

We next looked at cellular localization of eIF2B and YBR159W using strains with subunits of eIF2B endogenously tagged with GFP and the YBR159W-RFP expression plasmid YCp-YBR159W-dsRed. To show that the RFP-tagged YBR159W allele was functional, the plasmid YCp-YBR159W-dsRed complemented the *ybr159w*Δ null strain AL401 (data not shown). Confocal microscopy of the dual-fluorescentce-labeled strains was used to look for colocalization between eIF2B and YBR159W (Fig. 8). As observed in previous studies and Fig. 2, YBR159W localized to membranes corresponding to the ER (Fig. 8). Using strains with different eIF2B subunits tagged with GFP, we observed eIF2B localized as one to two large foci (Fig. 8). In addition, GFP-tagged eIF2B is seen dispersed throughout the cytoplasm (data not shown). Sur-

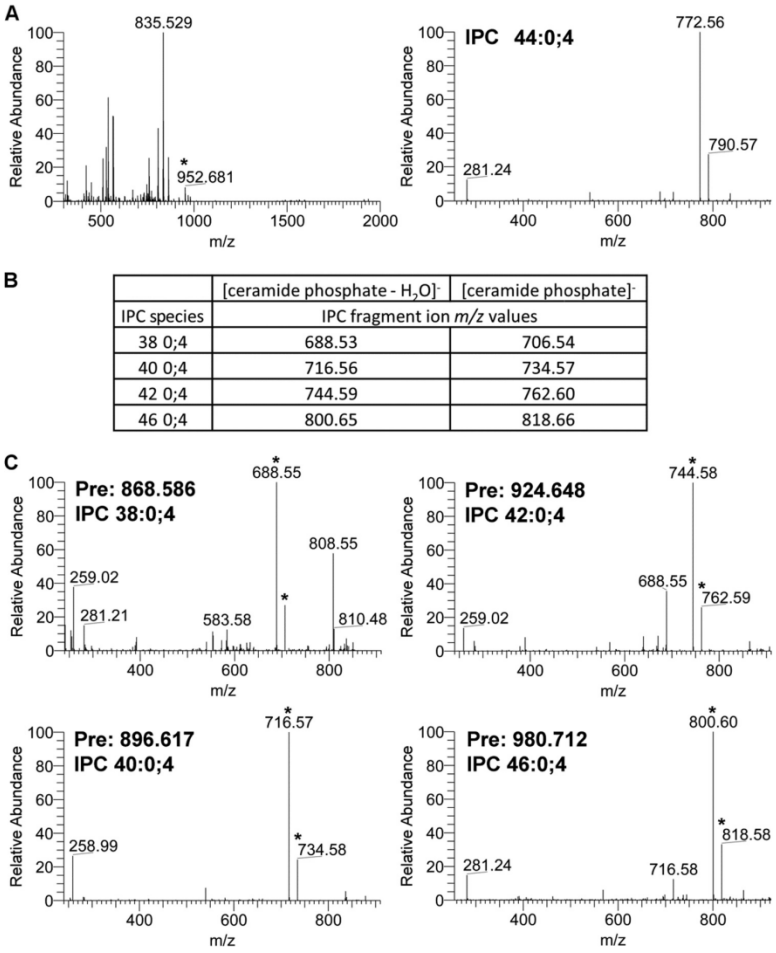



**FIG 6** Validation and identification of IPCs. (A) MS precursor and MS/MS fragmentation spectrum for IPC 44:0;4 from WT yeast. The observed precursor ion *m/z* 952.681 represents the expected ion IPC 44:0;4 (952.686). The observed *m/z* 835.529 corresponds to the two PIs, PI 16:0-18:1 and PI 16:1-18:0, used to normalize the relative abundance of each IPC species. In the lower MS/MS spectra of the 952.681 precursor ion, the first and second most abundant peaks correspond to the expected IPC 44:0;4 fragment ions [ceramide phosphate-$H_2O$]$^-$ at *m/z* 772.62 and [ceramide phosphate]$^-$ at *m/z* 790.63. (B) Theoretical fragmentation database for IPCs 38:0;4, 40:0;4, 42:0;4, and 46:0;4. Shown are the theoretical *m/z* values for fragment ions [ceramide phosphate-$H_2O$]$^-$ and [ceramide phosphate]$^-$ for IPCs 38:0;4, 40:0;4, 42:0;4, and 46:0;4. (C) MS/MS fragmentation spectra for IPCs 38:0;4, 40:0;4, 42:0;4, and 46:0;4. The observed precursor *m/z* values "Pre" of 868.586, 896.617, 924.648, and 980.712 correspond to the expected *m/z* values of IPC 38:0;4 (868.592), IPC 40:0;4 (896.623), IPC 42:0;4 (924.655), and IPC 46:0;4 (980.717), respectively. In the MS/MS spectra, the peaks corresponding to the expected theoretical IPC fragment ions [ceramide phosphate-$H_2O$]$^-$ and [ceramide phosphate]$^-$ are marked with an asterisk. In each case, the peak corresponding to the expected [ceramide phosphate-$H_2O$]$^-$ fragment ion was the most intense ion in the MS/MS spectrum.

prisingly, the confocal microscopy images did not convincingly show the majority of YBR159W signal colocalizing with eIF2B subunits. Because we observed that the 2B bodies localize near the ER membrane-bound YBR159W, we performed a statistical analysis to test if eIF2B and YBR159W colocalize. We examined 221 individual 2B bodies from 140 dually labeled cells by pooling results from the YCp-YBR159W-dsRed transformed GCD1-GFP, GCD6-GFP, and GCD7-GFP strains (AL405, AL406, and AL407). We found that 60.1% ± 6.6% of 2B bodies examined showed partial colocalization with a bright area of YBR159W signal. Based on the area of the cell taken up by bright areas of YBR159W signal, it would be expected that only 30.7% ± 6.9% of 2B bodies would colocalize with YBR159W signal if the two signals were independent of each other. Student *t* test ($P = 2.9 \times 10^{-8}$) shows this difference to be significant.

To observe the effects of the *ybr159wΔ* deletion on eIF2B localization, we performed live cell imaging using epifluorescence microscopy on the yeast strains *ybr159wΔ* GCD7-GFP (AL403),

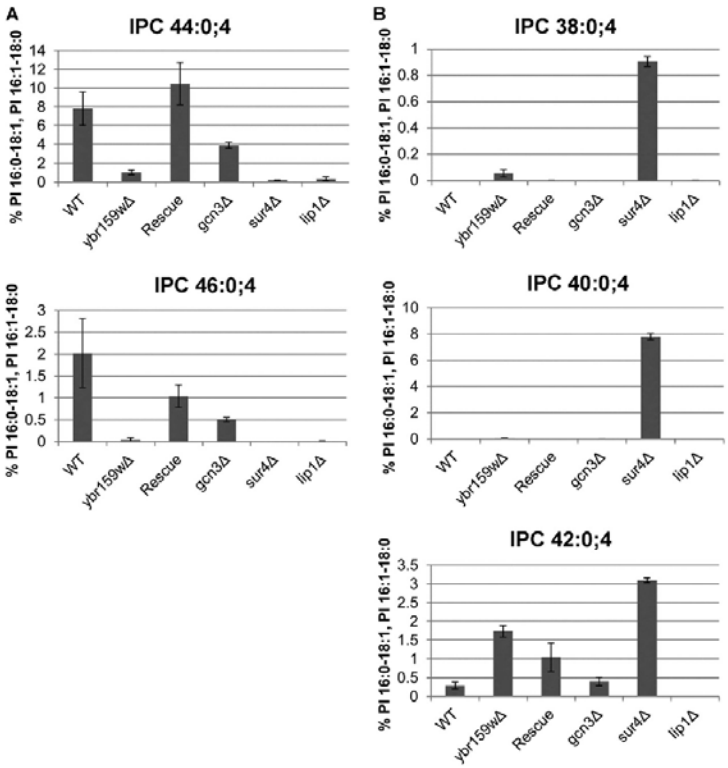



FIG 7 Fatty acid profiling of WT and mutant yeast strains. (A) Longer-chain fatty acid-containing sphingolipid species. IPC species with 44 and 46 carbon-containing acyl chains are shown. The VLCFA and ceramide mutants *sur4*Δ and *lip1*Δ synthase are included as controls. (B) Shorter-chain fatty acid-containing sphingolipid species. Three IPC species with 38, 40, and 42 carbon-containing acyl chains are shown. For both panels A and B, the data represent the percentage of signal of each lipid species normalized to the signals of the PI 16:0-18:1 and PI 16:1-18:0 ions.

GCD7-GFP (AL429), and *ybr159w*Δ GCD7-GFP, [YCp-YBR159W] (AL404). Cells from the GCD7-GFP control strain were found to contain one to two large 2B bodies (Fig. 9A). In the *ybr159w*Δ strain AL403, eIF2B appeared as multiple foci (Fig. 9A).

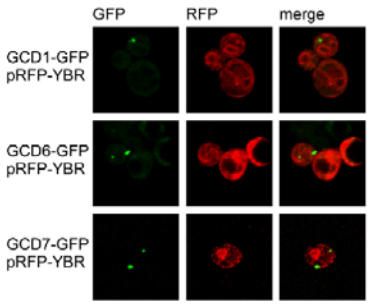



FIG 8 eIF2B and YBR159W localization in live cells. Confocal microscopy of live yeast cells showing localization of eIF2B subunits in relation to the localization of YBR159W. eIF2B subunits are endogenously tagged with GFP, while YBR159W-RFP is expressed on a centromeric plasmid.

The *ybr159w*Δ phenotype of AL403 was rescued by expression of plasmid YCp-YBR159W in strain AL404 (Fig. 9A). Using these strains, we counted the number of cells containing one to two large 2B bodies compared to the number of cells having the multiple eIF2B foci or diffuse cytoplasmic localization. We found that the majority of *ybr159w*Δ cells had a multiple eIF2B focus phenotype (Table 4). For the GCD7-GFP WT control strain, no cells had the multiple eIF2B focus phenotype and a majority of cells had either one or two 2B bodies. The rescued *ybr159w*Δ GCD7-GFP, [YCp-YBR159W] strain AL404 did not have multiple eIF2B foci (Table 4). To show that the 2B body phenotype was independent of the GFP-tagged alleles, we performed immunofluorescence microscopy on untagged yeast strains using polyclonal antibody against the eIF2B subunit GCD6 (Fig. 9B). We observed the one to two large 2B body focus phenotype for the majority of the WT control AL400 cells, while the majority of the *ybr159w*Δ cells (AL401) displayed multiple eIF2B foci (Table 4). The *ybr159w*Δ [YCp-YBR159W] rescue strain AL402 showed a majority of cells had eIF2B present as either a single 2B focus or no detectable foci (Table 4). The VLCFA and ceramide synthase mutants AL413 (*fen1*Δ), AL414 (*sur4*Δ), and RH5994 (*lip1*Δ) were all found to have the multiple eIF2B focus phenotype (Table 4).

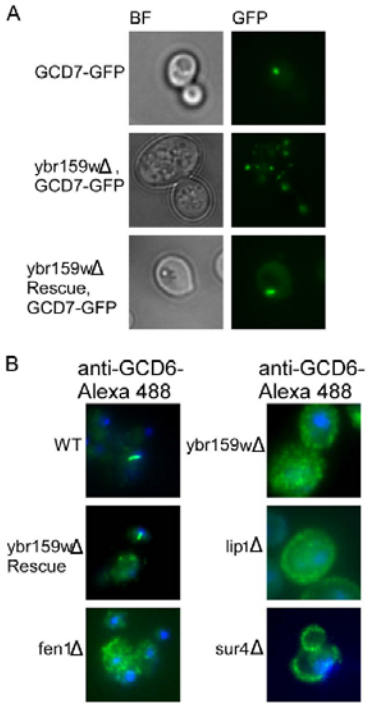



FIG 9 eIF2B localization in the *ybr159w*Δ background. (A) Live cell fluorescence microscopy of endogenously tagged eIF2B subunit GCD7-GFP. Bright-field (BF) images are included for clarity. (B) Immunofluorescence microscopy of formaldehyde-fixed yeast cells. A polyclonal antibody against yeast eIF2B subunit GCD6 was used along with an Alexa Fluor 488-tagged secondary. Nuclei are stained with DAPI for clarity.

Because eIF2B is thought to be a soluble cytoplasmic protein and YBR159W has been shown to be an integral membrane protein in the ER (26), we performed membrane float experiments to determine whether a population of eIF2B complexes physically interacted with lipid membranes. The lack of signal from the Western blot for the control yeast GAPDH analogs TDH1, TDH2, and TDH3 in the membrane fraction showed that the fractionation was efficient at separating cytoplasmic proteins from membrane-associated proteins (Fig. 10A). A significant lipid membrane signal was seen for the ER proteins YBR159W and DPM1. A portion of the YBR159W and control ER membrane protein DPM1 signals was still present in the soluble fraction, indicating that the membrane-associated proteins do not appear to completely separate from the soluble fraction. The membrane float experiments showed that in WT AL400 cells, a significant fraction of the eIF2B subunit GCD6 localized to the lipid membrane fractions (Fig. 10A). The Western blot profiles of the membrane and soluble fractions for GCD6 showed the same pattern as the known ER membrane proteins YBR159W and DPM1 (Fig. 10A). Interestingly, the SUI2 component of eIF2 also showed a similar membrane association pattern. The data indicate a fraction of eIF2 complexes are associated with membranes in yeast cells.

To validate our observation that eIF2B is membrane associated, we used whole-cell extracts prepared from TAP-tagged eIF2B and control strains and the membrane float separation experiment to collect fractions from the membrane-associated and soluble protein region of the density gradients. Next, we performed a modified TAP purification on each fraction and analyzed the affinity-purified complexes using LC-MS/MS. Our data showed that the interaction between eIF2B and YBR159W was still present in both the membrane-associated and soluble protein fractions (Fig. 10B). Because of the incomplete separation of membrane proteins in the assay, it is not known whether both soluble and membrane-associated eIF2Bs interact with YBR159W or whether only membrane-associated eIF2B interacts with YBR159W. To determine whether YBR159W was required for eIF2B's membrane association, we performed the membrane flotation assay and Western blot analyses using the *ybr159w*Δ strain AL401. We found that eIF2B associated with the membrane fraction in the *ybr159w*Δ strain at levels similar to those seen in the control strain (Fig. 10C). Overall, the membrane float experiments showed that a fraction of yeast eIF2B is associated with membranes but that the interaction is independent of YBR159W.

To determine whether the membrane association seen for eIF2B is possibly mediated by rough ER-bound ribosomes, we performed a subcellular fractionation experiment to isolate smooth membranes. Cell lysates from WT strain AL400 were treated with either elevated levels of EDTA or the ribosome-releasing antibiotic puromycin (63). After fractionation and Western blotting, ribosomal protein signal in the insoluble membrane fraction was significantly reduced in both the EDTA- and puromycin-treated cell extracts compared to untreated control extracts. However, the eIF2B signal in the rough or smooth membrane fraction did not noticeably change (Fig. 10D). The data indicate that the eIF2B-membrane association is independent of ribosomes.

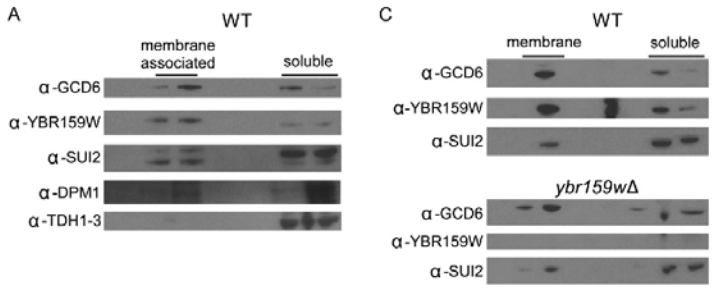
## DISCUSSION

Previous large-scale yeast interactions studies failed to show eIF2B interacting with the VLCFA pathway (64–66). We show using

TABLE 4 Statistics for eIF2B localization phenotypes in live yeast cells (group 1) and for eIF2B localization phenotypes via immunofluorescence analysis of fixed yeast cells (group 2)[a]

| Strain | No. of cells | Single 2B body (%) | Multiple foci (%) | No foci (%) |
|---|---|---|---|---|
| Group 1 | | | | |
| WT | 173 | 50.9 | 0.0 | 49.1 |
| *ybr159w*Δ mutant | 122 | 3.3 | 71.3 | 24.6 |
| Rescue | 72 | 44.4 | 0.0 | 55.6 |
| Group 2 | | | | |
| WT | 105 | 63.8 | 7.6 | 28.6 |
| *ybr159w*Δ mutant | 59 | 1.7 | 83.1 | 15.3 |
| Rescue | 111 | 35.1 | 15.3 | 49.5 |
| *lip1*Δ mutant | 29 | 6.9 | 93.1 | 0.0 |
| *fen1*Δ mutant | 96 | 4.2 | 63.5 | 32.3 |
| *sur4*Δ mutant | 122 | 1.6 | 58.2 | 40.2 |

[a] The group 1 strains are described in Fig. 9A. The group 2 strains are described in Fig. 9B.




B

| Gene | TAP-YBR Membrane | TAP-YBR-Soluble | TAP-GCD7 Membrane | TAP-GCD7 Soluble | TAP-GCD2 Membrane | TAP-GCD2 Soluble |
|---|---|---|---|---|---|---|
| YBR159W | 16(57.06) | 12(48.41) | 2(8.36) | 3(17.58) | 4(20.75) | 3(18.16) |
| GCD1 | 3(10.38) | 2(7.09) | 7(20.59) | 12(35.29) | 1(3.11) | 1(3.11) |
| GCD2 | 2(5.07) | 2(4.3) | 9(13.82) | 5(9.68) | 1(2.15) | 2.(4.3) |
| GCN3 | 1(7.54) | - | 5(24.26) | 6(34.43) | 1(5.9) | 2(16.07) |
| GCD6 | 8(16.57) | 5(12.92) | 13(27.81) | 11(21.63) | 6(14.47) | 2(4.21) |
| GCD7 | 1(2.89) | 1(7.35) | 7(15.49) | 6(22.83) | 3(11.55) | 1(2.89) |

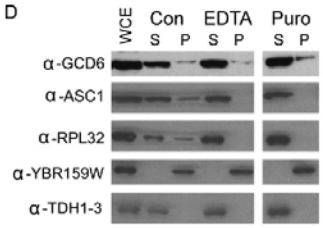



FIG 10 eIF2B and YBR159W localization using membrane flotation assays. (A) Western blot of membrane flotation assay fractions using protein extracts from WT yeast showing the localization of the eIF2B subunit GCD6 and YBR159W. Controls include the eIF2 subunit SUI2, the ER integral membrane protein DPM1, and the cytosolic protein GAPDH. The genes encoding TDH1 to TDH3 are the three GAPDH genes in yeast. The lanes represent 20% of fractions from the membrane flotation gradients. Labels show the location of the membrane-associated and soluble protein fractions. (B) MS analysis of affinity-purified TAP complexes from the membrane and soluble fractions of membrane flotation experiments. Unique peptides, the percent coverage, and "–" are all as described in Fig. 2A. (C) Western blot of membrane flotation assay fractions comparing WT and *ybr159w*Δ strains. The conditions are the same as in panel A. (D) Western blot of crude fractionation following EDTA or puromycin treatment. Abbreviations: WCE, whole-cell extract; Con, untreated control; EDTA, EDTA treatment; Puro, puromycin treatment; S, supernatant; P, pellet. ASC1 is a component of the small ribosomal subunit and RPL32 is a large ribosomal subunit protein. Lanes represent 15 μl of whole-cell extract (WCE) following fractionation, and pellets were resuspended in starting volume.

TAP-tagged and GFP-tagged affinity purifications, as well as yeast-two-hybrid analysis, that the VLCFA keto-reductase YBR159W interacts with the translation initiation factor complex eIF2B. Because our unpublished proteomic screen of translation factor interactions identified YBR159W interacting with eIF2B, we named the *S. cerevisiae* locus IFA38 for initiation factor-associated protein of 38 kDa (Link et al., unpublished). Affinity purification and LC-MS/MS experiments show that YBR159W copurifies with all five subunits of eIF2B and not in controls. No other member of the VLCFA pathway copurifies in the eIF2B affinity purifications. Interestingly, the TAP-tagged members of the VLCFA pathway do not seem to strongly interact with each other. Our Y2H data suggest that the eIF2B subunits GCD6 and GCD7 physically interact with YBR159W.

The interaction between the VLCFA synthesis pathway and the eIF2B translation initiation pathway presents a number of possibilities. Is one pathway regulating the other or vice versa? It can be hypothesized that the cell might need to regulate VLCFA synthesis if translation is disrupted. Alternatively, it might be advantageous to reduce translational activity if VLCFAs are being downregulated. Finally, the YBR159W-eIF2B complex could be involved in a novel function. A link between a translation initiation factor and lipid membranes is not totally unique. Experiments in human cells have shown an interaction between the translation initiation factor eIF4E and the Golgi apparatus (67).

To test the hypothesis that YBR159W and VLCFA synthesis play a role in translation, we used [$^{35}$S]methionine incorporation and polysome profiling to assay translation activity in mutant strains. Both experiments show a reduction in the translation rate for the *ybr159w*Δ strain. However, a similar phenotype is seen for the slow-growing *lip1*Δ strain. The VLFCA mutant *fen1*Δ and *sur4*Δ strains have WT growth rates and do not share a translation defect with the slower-growing members of the pathway. It is not known whether the cause of the translation

defect seen in the *ybr159wΔ* strain is directly related to its interaction with eIF2B or is an indirect consequence of slow growth or a VLCFA defect.

When GCN4 expression is examined using the GCN4-LacZ assay, the *ybr159wΔ* strain has WT levels of GCN4 induction. The GCN4-LacZ assay was normalized to protein concentration so the slow growth rate of the *ybr159wΔ* strain should not affect the results. The data indicate that the *ybr159wΔ* strain does not have a defect in the GCN4 pathway. We cannot rule out the possibility that the slow growth of the *ybr159wΔ* strain masks a subtle defect in eIF2B's GEF activity unrelated to the GCN4 pathway. Our affinity purification experiments of eIF2B in a *ybr159wΔ* deletion background showed that the eIF2B complex is intact. A Western blot of *ybr159wΔ* strains showed that the overall abundance of eIF2B was lower in the deletion background compared to WT. It is not clear whether the lower level of eIF2B is caused by the slow-growth phenotype of the *ybr159wΔ* null background or some other factor.

To test the hypothesis that eIF2B plays a role in VLCFA synthesis, several limitations arose that made answering the question problematic. Of the five yeast eIF2B subunits, only GCN3 is nonessential. The *gcn3Δ* strain did not show a defect in VLCFA production or utilization. Although the *gcn3Δ* strain showed a reduction in the sphingolipid species IPC 44:0;4 and IPC 46:0;4, it did not show a concomitant rise in shorter-chain fatty acid-containing IPC species indicative of a defect in VLCFA production. The presence of shorter-chain sphingolipids would indicate the cell is trying to compensate for a lack of VLCFAs. Therefore, we postulate the lower levels of IPC 44:0;4 and IPC 46:0;4 seen in the *gcn3Δ* strain are unrelated to a defect in VLCFA production. The VLCFA defect in the *ybr159wΔ* strain is modest, with only a small increase in the shorter-chain fatty acid-containing sphingolipids. The loss of IPC 46:0;4 is the strain's most striking characteristic. Previous work suggests that Ayr1p is able to perform 3-ketoacyl activity in the absence of YBR159W (30). The same study showed that *ayr1* and *ybr159w* are synthetically lethal (30).

A *gcn3Δ* null strain is unable to fully derepress GCN4 expression during amino acid starvation (68). GCN4 is a transcription factor involved in the expression of several hundred genes during a wide variety of cellular stresses (69). Although growth conditions for the *gcn3Δ* strain should not have activated a stress response, we suspected analysis of the lipid content of the *gcn3Δ* strain could prove problematic if the VLCFA pathway was a downstream target of the GCN4 transcription factor. We examined the effects of loss of GCN4 using expression data for *gcn4Δ* strains from the Gene Expression Omnibus (GEO) Database (70). Two separate data sets showed no significant changes in the expression of various VLCFA genes (data not shown; GEO accession no. GSE24057 [71] and GSE25582). We concluded that under the conditions used for the analysis of sphingolipids, the loss of GCN4 did not significantly alter VLCFA gene expression. We concluded that our *gcn3Δ* strain was not experiencing alterations in VLCFA gene expression due to the repression of GCN4. The lack of a direct translation defect in the *ybr159wΔ* strain and the lack of a VLCFA defect in the *gcn3Δ* strain suggest that there is no significant cross talk between the GEF and VLCFA pathways.

Membrane flotation and subcellular fractionation assays show eIF2B interacts with lipid membranes. Our data and previous studies showed YBR159W is an integral membrane protein that colocalizes with the ER membrane (26, 27). We interpret these findings to mean that the membranes eIF2B is interacting with are ER membranes. It is unknown whether ER-associated eIF2B is actively engaged in guanine nucleotide exchange. A number of conclusions can be made about this ER membrane-interacting eIF2B. First, the eIF2B-membrane interaction is not mediated by rough ER-bound ribosomes. Treatment of cell extracts with EDTA or puromycin greatly reduces the amount of ribosomes that fractionate with lipid membranes but does not reduce the portion of eIF2B that fractionates with membranes. This fits the prevailing theory that eIF2B's role in translation is independent of the ribosome (72). Second, YBR159W is not required for the interaction. The *ybr159wΔ* null strain does not affect eIF2B's interaction with the membrane showing that the interaction of eIF2B with ER membranes is YBR159W independent. This indicates that another factor(s) is possibly required.

Confocal microscopy shows that the majority of 2B bodies are in close proximity to YBR159Wp and ER membranes, supporting the model that 2B bodies and the ER interact. This could be taken to indicate that the eIF2B shown to interact with ER membranes resides in 2B bodies. A possible conflicting interpretation of the data is that YBR159W-RFP is being overexpressed and its localization is an artifact. The colocalization experiment used a RFP-tagged YBR159W allele expressed from a GPD promoter on a centromeric plasmid. Global protein expression analysis shows that the GPD promoter's target protein, TDH3, is expressed at roughly four times that of YBR159W (40). The fact that the RFP-tagged YBR159W localization agrees with endogenously expressed YBR159W-GFP localization leads us to believe that artifacts caused by the RFP tagged construct are not disrupting YBR159W's localization. In addition, the RFP-tagged allele complements a *ybr159wΔ* null strain. How and why eIF2B might be interacting with the ER membrane is unknown. A population of membrane-interacting 2B bodies might explain recent findings that 2B bodies can exist in a mobile or static state with mobile 2B bodies free in the cytoplasm and static 2B bodies being associated with membranes (73). Further work is needed to prove this hypothesis.

The observation that the *ybr159wΔ* null strain leads to multiple eIF2B foci is intriguing. This phenotype is also seen in other VLCFA mutants. The fact that these mutants all display disrupted lipid membranes lends itself to the theory that properly formed membranes are required for the integrity of 2B bodies. An intriguing question is whether the membrane disruption prevents the 2B bodies from forming properly or whether the 2B bodies are unable to be maintained once formed? For the first model, an as-yet-unknown factor in membranes required for 2B body formation could be disrupted and cause 2B bodies to form throughout the cell. We speculate that this membrane-associated factor could serve as a nucleating site for the formation of 2B bodies. The second model would predict that membrane disruption is affecting a factor needed for 2B body stability. Loss of this factor leads to 2B bodies dissociating into multiple smaller foci. A previous study showed VLCFAs were important for lipid raft formation (25). It is possible that lipid raft disruption in the VLCFA mutants causes the multiple eIF2B focus phenotype. Translation assays using the *ybr159wΔ* strain suggested the disruption of 2B bodies into multiple foci does not affect translation. The translation activity of yeast cells does not appear to be affected by the change from a single 2B body to multiple eIF2B foci.

Our work sheds light on the recently discovered 2B body. The

data show a relationship between eIF2B localization and an ER membrane-bound protein. We discovered eIF2B's membrane colocalization while examining its interaction with YBR159W. Our data show that YBR159W is not necessary for 2B's colocalization to the membrane. The primary mediator of eIF2B's membrane association is unknown. It remains to be determined whether the translation defect seen in the *ybr159wΔ* strain is the cause of the strain's slow growth or vice versa. Further experiments are required to determine the functional role of YBR159W interacting with eIF2B.

## ACKNOWLEDGMENTS

C.M.B. was supported by National Institutes of Health (NIH) training grant T32 AI007611, and P.S. and A.J.L. were supported by NIH grant GM64779. The experiments, data analysis, and presentation of fluorescence microscope images were performed in part through the use of the VUMC Cell Imaging Shared Resource, which is supported by NIH grants CA68485, DK20593, DK58404, HD15052, and DK59637.

We thank Tom Dever, Alan Hinnebusch, Howard Riezman, and Jonathan Warner for reagents and yeast strains.

## REFERENCES

1. Preiss T, Hentze WM. 2003. Starting the protein synthesis machine: eukaryotic translation initiation. Bioessays 25:1201–1211.
2. Sonenberg NH, Mathews JM. 2000. Translational control of gene expression. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. Pavitt GD. 2005. eIF2B, a mediator of general and gene-specific translational control. Biochem. Soc. Trans. 33:1487–1492.
4. Rowlands AG, Panniers R, Henshaw EC. 1988. The catalytic mechanism of guanine nucleotide exchange factor action and competitive inhibition by phosphorylated eukaryotic initiation factor 2. J. Biol. Chem. 263:5526–5533.
5. Zhan K, Narasimhan J, Wek RC. 2004. Differential activation of eIF2 kinases in response to cellular stresses in *Schizosaccharomyces pombe*. Genetics 168:1867–1875.
6. Schneider RJ, Mohr I. 2003. Translation initiation and viral tricks. Trends Biochem. Sci. 28:130–136.
7. Harding HP, Zhang Y, Ron D. 1999. Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase. Nature 397:271–274.
8. Wek RC, Ramirez M, Jackson BM, Hinnebusch AG. 1990. Identification of positive-acting domains in GCN2 protein kinase required for translational activation of GCN4 expression. Mol. Cell. Biol. 10:2820–2831.
9. Olsen DS, Jordan B, Chen D, Wek RC, Cavener DR. 1998. Isolation of the gene encoding the *Drosophila melanogaster* homolog of the *Saccharomyces cerevisiae* GCN2 eIF-2α kinase. Genetics 149:1495–1509.
10. Sood R, Porter AC, Olsen DA, Cavener DR, Wek RC. 2000. A mammalian homologue of GCN2 protein kinase important for translational control by phosphorylation of eukaryotic initiation factor-2α. Genetics 154:787–801.
11. Fabian JR, Kimball SR, Heinzinger NK, Jefferson LS. 1997. Subunit assembly and guanine nucleotide exchange activity of eukaryotic initiation factor-2B expressed in Sf9 cells. J. Biol. Chem. 272:12359–12365.
12. Gomez E, Pavitt GD. 2000. Identification of domains and residues within the epsilon subunit of eukaryotic translation initiation factor 2B (eIF2Bε) required for guanine nucleotide exchange reveals a novel activation function promoted by eIF2B complex formation. Mol. Cell. Biol. 20:3965–3976.
13. Pavitt GD, Ramaiah KV, Kimball SR, Hinnebusch AG. 1998. eIF2 independently binds two distinct eIF2B subcomplexes that catalyze and regulate guanine-nucleotide exchange. Genes Dev. 12:514–526.
14. Bushman JL, Asuru AI, Matts RL, Hinnebusch AG. 1993. Evidence that GCD6 and GCD7, translational regulators of GCN4, are subunits of the guanine nucleotide exchange factor for eIF-2 in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 13:1920–1932.
15. Pavitt GD, Yang W, Hinnebusch AG. 1997. Homologous segments in three subunits of the guanine nucleotide exchange factor eIF2B mediate translational regulation by phosphorylation of eIF2. Mol. Cell. Biol. 17:1298–1313.
16. Kubica N, Jefferson LS, Kimball SR. 2006. Eukaryotic initiation factor 2B and its role in alterations in mRNA translation that occur under a number of pathophysiological and physiological conditions. Prog. Nucleic Acid Res. Mol. Biol. 81:271–296.
17. Hinnebusch AG. 1985. A hierarchy of *trans*-acting factors modulates translation of an activator of amino acid biosynthetic genes in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 5:2349–2360.
18. Campbell SG, Hoyle NP, Ashe MP. 2005. Dynamic cycling of eIF2 through a large eIF2B-containing cytoplasmic body: implications for translation control. J. Cell Biol. 170:925–934.
19. Campbell SG, Ashe MP. 2006. Localization of the translational guanine nucleotide exchange factor eIF2B: a common theme for GEFs? Cell Cycle 5:678–680.
20. Rossler H, Rieck C, Delong T, Hoja U, Schweizer E. 2003. Functional differentiation and selective inactivation of multiple *Saccharomyces cerevisiae* genes involved in very-long-chain fatty acid synthesis. Mol. Genet. Genomics 269:290–298.
21. Stoops JK, Wakil SJ. 1978. The isolation of the two subunits of yeast fatty acid synthetase. Biochem. Biophys. Res. Commun. 84:225–231.
22. Welch JW, Burlingame AL. 1973. Very long-chain fatty acids in yeast. J. Bacteriol. 115:464–466.
23. Dittrich F, Zajonc D, Huhne K, Hoja U, Ekici A, Greiner E, Klein H, Hofmann J, Bessoule JJ, Sperling P, Schweizer E. 1998. Fatty acid elongation in yeast: biochemical characteristics of the enzyme system and isolation of elongation-defective mutants. Eur. J. Biochem. 252:477–485.
24. Dickson RC, Sumanasekera C, Lester RL. 2006. Functions and metabolism of sphingolipids in *Saccharomyces cerevisiae*. Prog. Lipid Res. 45:447–465.
25. Gaigg B, Toulmay A, Schneiter R. 2006. Very long-chain fatty acid-containing lipids rather than sphingolipids per se are required for raft association and stable surface transport of newly synthesized plasma membrane ATPase in yeast. J. Biol. Chem. 281:34135–34145.
26. Klein HP. 1957. Some observations on a cell free lipid synthesizing system from *Saccharomyces cerevisiae*. J. Bacteriol. 73:530–543.
27. Abraham S, Chaikoff IL, Bortz WM, Klein HP, Den H. 1961. Particle involvement in fatty acid synthesis in liver and yeast systems. Nature 192:1287–1288.
28. Tehlivets O, Scheuringer K, Kohlwein SD. 2007. Fatty acid synthesis and elongation in yeast. Biochim. Biophys. Acta 1771:255–270.
29. Beaudoin F, Gable K, Sayanova O, Dunn T, Napier JA. 2002. A *Saccharomyces cerevisiae* gene required for heterologous fatty acid elongase activity encodes a microsomal beta-keto-reductase. J. Biol. Chem. 277:11481–11488.
30. Han G, Gable K, Kohlwein SD, Beaudoin F, Napier JA, Dunn TM. 2002. The *Saccharomyces cerevisiae YBR159w* gene encodes the 3-ketoreductase of the microsomal fatty acid elongase. J. Biol. Chem. 277:35440–35449.
31. Oh CS, Toke DA, Mandala S, Martin CE. 1997. ELO2 and ELO3, homologues of the *Saccharomyces cerevisiae ELO1* gene, function in fatty acid elongation and are required for sphingolipid formation. J. Biol. Chem. 272:17376–17384.
32. Tuller G, Prein B, Jandrositz A, Daum G, Kohlwein SD. 1999. Deletion of six open reading frames from the left arm of chromosome IV of *Saccharomyces cerevisiae*. Yeast 15:1275–1285.
33. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ. 2005. Exploration of the function and organization of the yeast early secretory pathway through an epistatic mini-array profile. Cell 123:507–519.
34. Vallee B, Riezman H. 2005. Lip1p: a novel subunit of acyl-CoA ceramide synthase. EMBO J. 24:730–741.
35. Dickson RC. 2008. Thematic review series: sphingolipids: new insights into sphingolipid metabolism and function in budding yeast. J. Lipid Res. 49:909–921.
36. Preuss D, Mulholland J, Kaiser CA, Orlean P, Albright C, Rose MD, Robbins PW, Botstein D. 1991. Structure of the yeast endoplasmic reticulum: localization of ER proteins using immunofluorescence and immunoelectron microscopy. Yeast 7:891–911.
37. Lowe M, Barr FA. 2007. Inheritance and biogenesis of organelles in the secretory pathway. Nat. Rev. Mol. Cell Biol. 8:429–439.
38. David Amberg C, DJB, Jeffrey Strathern N. 2005. Methods in yeast genetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
39. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM,

Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW. 1999. Functional characterization of the *Saccharomyces cerevisiae* genome by gene deletion and parallel analysis. Science 285:901–906.

40. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature 425:737–741.
41. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. Nature 425:686–691.
42. James P, Halladay J, Craig EA. 1996. Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. Genetics 144: 1425–1436.
43. Hudson JR, JR, Dawson EP, Rushing KL, Jackson CH, Lockshon D, Conover D, Lanciault C, Harris JR, Simmons SJ, Rothstein R, Fields S. 1997. The complete set of predicted genes from *Saccharomyces cerevisiae* in a readily usable form. Genome Res. 7:1169–1173.
44. Alberti S, Gitler AD, Lindquist S. 2007. A suite of Gateway cloning vectors for high-throughput genetic analysis in *Saccharomyces cerevisiae*. Yeast 24:913–919.
45. Powell DW, Weaver CM, Jennings JL, McAfee KJ, He Y, Weil PA, Link AJ. 2004. Cluster analysis of mass spectrometry data reveals a novel component of SAGA. Mol. Cell. Biol. 24:7249–7259.
46. Sanders SL, Jennings J, Canutescu A, Link AJ, Weil PA. 2002. Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. Mol. Cell. Biol. 22:4723–4738.
47. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR, III. 1999. Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. 17:676–682.
48. Link AJ, Fleischer TC, Weaver CM, Gerbasi VR, Jennings JL. 2005. Purifying protein complexes for mass spectrometry: applications to protein translation. Methods 35:274–290.
49. McAfee KJ, Duncan DT, Assink M, Link AJ. 2006. Analyzing proteomes and protein function using graphical comparative analysis of tandem mass spectrometry results. Mol. Cell. Proteomics 5:1497–1513.
50. Ejsing CS, Sampaio JL, Surendranath V, Duchoslav E, Ekroos K, Klemm RW, Simons K, Shevchenko A. 2009. Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 106:2136–2141.
51. Ejsing CS, Moehring T, Bahr U, Duchoslav E, Karas M, Simons K, Shevchenko A. 2006. Collision-induced dissociation pathways of yeast sphingolipids and their molecular profiling in total lipid extracts: a study by quadrupole TOF and linear ion trap-orbitrap mass spectrometry. J. Mass Spectrom. 41:372–389.
52. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Jr, Murphy RC, Raetz CR, Russell DW, Subramaniam S. 2007. LMSD: LIPID MAPS structure database. Nucleic Acids Res. 35:D527–D532.
53. Fields S, Song O. 1989. A novel genetic system to detect protein-protein interactions. Nature 340:245–246.
54. Bergmann JE, Fusco PJ. 1988. The M protein of vesicular stomatitis virus associates specifically with the basolateral membranes of polarized epithelial cells independently of the G protein. J. Cell Biol. 107:1707–1715.
55. Rose M, Botstein D. 1983. Construction and use of gene fusions to *lacZ* (β-galactosidase) that are expressed in yeast. Methods Enzymol. 101:167–180.
56. Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH image to ImageJ: 25 years of image analysis. Nat. Methods 9:671–675.
57. Terasaki M, Song J, Wong JR, Weiss MJ, Chen LB. 1984. Localization of endoplasmic reticulum in living and glutaraldehyde-fixed cells with fluorescent dyes. Cell 38:101–108.
58. Gerbasi VR, Weaver CM, Hill S, Friedman DB, Link AJ. 2004. Yeast Asc1p and mammalian RACK1 are functionally orthologous core 40S ribosomal proteins that repress gene expression. Mol. Cell. Biol. 24:8276–8287.
59. Orlean P. 1990. Dolichol phosphate mannose synthase is required in vivo for glycosyl phosphatidylinositol membrane anchoring, O mannosylation, and N glycosylation of protein in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 10:5796–5805.
60. Orlean P, Albright C, Robbins PW. 1988. Cloning and sequencing of the yeast gene for dolichol phosphate mannose synthase, an essential protein. J. Biol. Chem. 263:17499–17507.
61. Schneiter R, Brugger B, Amann CM, Prestwich GD, Epand RF, Zellnig G, Wieland FT, Epand RM. 2004. Identification and biophysical characterization of a very-long-chain-fatty-acid-substituted phosphatidylinositol in yeast subcellular membranes. Biochem. J. 381:941–949.
62. Hinnebusch AG. 1994. Translational control of GCN4: an in vivo barometer of initiation-factor activity. Trends Biochem. Sci. 19:409–414.
63. Adelman MR, Sabatini DD, Blobel G. 1973. Ribosome-membrane interaction: nondestructive disassembly of rat liver rough microsomes into ribosomal and membranous components. J. Cell Biol. 56:206–229.
64. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147.
65. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St. Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440: 637–643.
66. Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, Fields S. 2005. Large-scale identification of yeast integral membrane protein interactions. Proc. Natl. Acad. Sci. U. S. A. 102:12123–12128.
67. Willett M, Brocard M, Davide A, Morley SJ. 2011. Translation initiation factors and active sites of protein synthesis colocalize at the leading edge of migrating fibroblasts. Biochem. J. 438:217–227.
68. Hannig EM, Hinnebusch AG. 1988. Molecular analysis of GCN3, a translational activator of GCN4: evidence for posttranslational control of GCN3 regulatory function. Mol. Cell. Biol. 8:4808–4820.
69. Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ. 2001. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol. Cell. Biol. 21:4347–4368.
70. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A. 2011. NCBI GEO: archive for functional genomics data sets: 10 years on. Nucleic Acids Res. 39:D1005–D1010.
71. Fendt SM, Oliveira AP, Christen S, Picotti P, Dechant RC, Sauer U. 2010. Unraveling condition-dependent networks of transcription factors that control metabolic pathway activity in yeast. Mol. Systems Biol. 6:432.
72. Merrick WC, Hershey JWB. 1996. The pathway and mechanism of eukaryotic protein synthesis, p 31–69. *In* Hershey JWB, Mathews MB, Sonenberg N (ed), Translation control. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
73. Taylor EJ, Campbell SG, Griffiths CD, Reid PJ, Slaven JW, Harrison RJ, Sims PF, Pavitt GD, Delneri D, Ashe MP. 2010. Fusel alcohols regulate translation initiation by inhibiting eIF2B to reduce ternary complex in a mechanism that may involve altering the integrity and dynamics of the eIF2B body. Mol. Biol. Cell 21:2202–2216.

# Appendix AB – Manuscript – 7: Saccharomyces cerevisiae Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein *ZNF9*

Biochemical and Biophysical Research Communications xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

# *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9

Morgan A. Sammons [a], Parimal Samir [b], Andrew J. Link [c,*]

[a] Department of Biological Sciences, Vanderbilt University, Nashville, TN 37232, USA
[b] Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA
[c] Department of Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

ARTICLE INFO

*Article history:*
Received 3 January 2011
Available online xxxx

*Keywords:*
Myotonic dystrophy type 2
DM2
ZNF9
GIS2
Yeast
IRES
Orthologue
Ribosome
Cap-independent translation

ABSTRACT

The myotonic dystrophy type 2 protein ZNF9/CNBP is a small nucleic acid binding protein proposed to act as a regulator of transcription and translation. The precise functions and activity of this protein are poorly understood. Previous studies suggested that ZNF9 regulates translation and facilitates the process of cap-independent translation through interactions with mRNA and the translating ribosome. To help determine the role played by ZNF9 in the activation of translation initiation, we combined genetic and biochemical analysis of the putative *ZNF9* ortholog *GIS2*, in the budding yeast *Saccharomyces cerevisiae*. Purification of the Gis2p protein followed by mass spectrometry based-proteomic analysis identified a large number of co-purifying ribosomal subunits and translation factors, strongly suggesting that Gis2p interacts with the protein translation machinery. Polysome profiling and ribosome isolation experiments confirm that Gis2p physically interacts with the translating ribosome. Interestingly, expression of yeast Gis2p in HEK293T cells activates cap-independent translation driven by the 5′UTR of the ODC gene. These data suggest that Gis2 is functionally orthologous to ZNF9 and acts as a cap-independent translation factor.



## 1. Introduction

The human disease myotonic dystrophy type 2 (DM2) is caused by a CCTG tetranucleotide repeat expansion in the first intron of the *znf9* gene, which codes for the protein ZNF9/CNBP [1]. ZNF9/CNPB is a small nucleic acid binding protein. Previous studies have implicated ZNF9 in a wide variety of molecular functions, ranging from the regulation of transcription to control of cell growth and proliferation [2–6]. Much of the ambiguity in the reported function of ZNF9 comes from the complexity of studying essential genes in mammalian model systems. A homozygous deletion of ZNF9/CNBP results in embryonic lethality in mice due to defects in brain development [3]. Reduced expression of ZNF9 results in truncated forebrains in chickens and severe craniofacial defects in zebrafish [4,5]. While these phenotypes provide evidence for the importance of the ZNF9 protein in growth and development, they do not illuminate the *in vivo* cellular functions and interactions of the protein.

Recently, we discovered that ZNF9 functions as a regulator of cap-independent or internal ribosome entry site (IRES)-mediated translation [7,8]. IRES-mediated translation in eukaryotes involves a number of factors, including IRES-specific *trans*-activating factors (ITAFs) that are thought to stabilize certain mRNA structures and facilitate their interactions with the ribosome [9–11]. ZNF9 directly binds to the IRES sequence in the 5′UTR of the ornithine decarboxylase (ODC) mRNA and facilitates the translation of this mRNA independent of the 5′-cap complex [8]. Other groups have also observed that ZNF9 acts as a regulator of translation [2,12–15], but the full scope of ZNF9's cellular interactions and any other processes regulated by ZNF9 are not well understood.

The translation machinery of the baker's yeast *Saccharomyces cerevisiae* and other single-celled eukaryotes is well conserved. The *S. cerevisiae* genome encodes one putative homolog of mammalian *ZNF9* called *GIS2*. *GIS2* was initially cloned as a multi-copy suppressor of the Gal-phenotype in a *snf1/mig1/srb8* yeast mutant [16]. The physiological role for *GIS2* in *S. cerevisiae* is unknown.

Because ZNF9 is absolutely required for the viability of vertebrate organisms [3,17], facile genetic analysis of ZNF9 function is limited in these complex animals. To address these issues, we have begun characterization of the putative ZNF9 homolog in *S. cerevisiae*. While Gis2p contains significant sequence similarity with the *Homo sapiens* ZNF9 protein, it is unknown whether the two proteins can be considered functional orthologs.

* Corresponding author. Address: Department of Microbiology and Immunology, Vanderbilt University School of Medicine, 1161 21st Ave. South, Nashville, TN 37232, USA. Fax: +1 615 343 7392.
*E-mail address:* andrew.link@vanderbilt.edu (A.J. Link).

0006-291X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2011.01.086

Please cite this article in press as: M.A. Sammons et al., *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9, Biochem. Biophys. Res. Commun. (2011), doi:10.1016/j.bbrc.2011.01.086

Using a series of biochemical and genetic assays, we demonstrate that Gis2p and ZNF9 are likely functional orthologs. Like ZNF9, Gis2p associates with translating ribosomes and copurifies with many ribosomal protein subunits as determined by mass spectrometry-based proteomics. Gis2p interaction with the ribosome is RNAse-sensitive, suggesting a mechanism for the observed ribosomal interactions. Expression of Gis2p in human cells is able to activate cap-independent translation of the ODC IRES. These data together suggest that the *S. cerevisiae* protein Gis2p and the mammalian protein ZNF9 are functional orthologs and provide a novel system in which to study the molecular functions of ZNF9 in translation and other essential cellular processes.

## 2. Materials and methods

### 2.1. Yeast strains and plasmids

Yeast genetic manipulations and media preparation were performed essentially as described [18]. The Δ*gis2* deletion strain used in this study had the entire Gis2 ORF replaced by a kanamycin cassette in the BY4743 background [19]. Knockout strains of Δ*gis2* were confirmed by PCR-based amplification across the Δ*gis2*::Kan$^R$ locus and phenotypically by growth on G418-containing medium. The Gis2-TAP yeast strain for purification of Gis2p has been previously described [20].

### 2.2. Tandem affinity purification and LC/MS/MS analysis

Gis2p was purified from 1 L cultures of the Gis2-TAP yeast strain grown to early stationary phase ($O.D._{600}$ 2–4). Gis2p and its associated proteins were isolated using a dual affinity protocol as previously described [21]. A 10% fraction of the eluted proteins was analyzed by SDS–PAGE on 4–12% Novex Bis-TRIS gels and silver stained. The remaining eluted proteins were reduced, alkylated, and digested with sequencing grade trypsin as previously described [21]. Trypsin-digested peptides were identified using two-dimensional microcapillary liquid chromatography coupled with an LTQ linear ion trap mass spectrometer as described previously [21]. The acquired mass spectra were correlated to a translated *S. cerevisiae* ORF database using Sequest [22]. The resulting peptide hits were processed, assembled, and analyzed using bioinformatics graphical comparative analysis tools (BIGCAT) as previously described [7,23,24].

### 2.3. Polysome profiling and ribosome salt wash experiments

Polysome profiling experiments were performed as previously described [21]. Ribosome pelleting was accomplished by centrifugation of whole cell lysates through a 1 M sucrose cushion (±10 mg/mL RNAse H or 50 mM EDTA) at 100,000*g* for 2 h in a TLA120.2 rotor. The resulting ribosome pellets were resuspended in ribosome solubilization buffer (20 mM TRIS–HCL (pH 7.5), 5 mM $MgCl_2$, and 6 M urea). Supernatant fractions were precipitated with TCA, washed with acetone, and resuspended in ribosome solubilization buffer. For ribosome salt wash analysis, ribosome pellets were isolated as described above. Pellets were washed in ice-cold PBS and resuspended in ribosome salt wash buffer containing 20 mM TRIS–HCl (pH 7.5), 5 mM $MgCl_2$, 1 mM β-mercaptoethanol, and either 100 mM, 250 mM, 500 mM, or 1 M potassium acetate. Resuspended ribosomes were then centrifuged at 100,000*g* for 2 h in a TLA120.2 rotor. Supernatant fractions were removed, precipitated with TCA, and analyzed by SDS–PAGE and western blotting.

### 2.4. Cell culture assay for cap-independent translation

HEK293T cells were cultured in 10 cm tissue culture-treated plates using DMEM and 10% fetal bovine serum at 37 °C and 5% $CO_2$. pcDNA3.1-hODC-IRES, pcDNA3.1-V5-LacZ, and pcDNA3.1-V5-ZNF9 have been previously described [8]. pcDNA3.1-V5-Gis2 was created by Gateway-mediated recombination of pcDNA3.1-V5-*ccdB* and pENTR-Gis2, which was created by cloning the *GIS2* coding sequence amplified from yeast genomic DNA using the primers 5′-CACCAT GTCTCAAAAAGCTTGTTACG-3′ and 5′-CTAAGCCTTTGGACAATCCT-3′. Cap-independent translation activity was assayed as previously described [8].

### 2.5. Cloning of *znf9* deletion mutants

To create deletion mutants of ZNF9, PCR products amplified using the following primers on full-length pcDNA3.1-V5-ZNF9 were cloned into pENTR-D/TOPO (Invitrogen). Antisense primer for the Δ1 and ΔRGG mutants: 5′-AGGCTGTAGCCTCAATTGTGCATTC-3′.

Sense primer for Δ1: 5′-CACCATGACTGGTGGAGGCCGTGGTCG-3′. Sense primer for ΔRGG: 5′-CACCATGATTTGTTATCGCTGTGGTGA-3′. Sense primer for the Δ7, Δ6, and Δ5 mutants: 5′-CACCATGAGCAGCAATGAGTGCTT-3′. Antisense primer for Δ7: 5′-TTATTCACTTGTCTTGCTGCAGT-3′. Antisense primer for Δ6: 5′-TTATTTGGTGCAGTCTTTTTGAA-3′. Antisense primer for Δ5: 5′-TTACTCATCTGCATGGTCGCAGT-3′. Destination clones were created by Gateway-mediated recombination of individual entry clones with pcDNA3.1-V5-*ccdB* (Invitrogen).

## 3. Results

### 3.1. S. cerevisiae Gis2p copurifies with the eukaryotic ribosome

Phylogenetic and sequence alignment analysis using the ClustalW algorithm suggests ZNF9/CNBP sequences are conserved in all eukaryotic organisms including *S. cerevisiae* (Fig. 1A and B) [25]. Human ZNF9/CNBP shares 36% sequence identify and 59% similarity with yeast Gis2p. Primary sequence analysis of *ZNF9/CNBP* and yeast Gis2p revealed the presence of seven CCHC zinc finger sequences (Fig. 1C) [26]. While an RGG box motif is identified in higher eukaryotic organisms, the RGG motif is noticeably absent from the *S. cerevisiaie* Gis2p protein. We sought to determine whether the high level of sequence similarity translated into functionally similar protein activities.

Previously, ZNF9 was found to associate with the translating ribosome and act as a regulator of cap-independent translation [7,8]. To examine whether Gis2p and ZNF9 have conserved interactions with the ribosome, we used tandem affinity purification followed by tandem mass spectrometry to identify proteins that associate with Gis2p. This approach has been extensively used to identify and characterize protein:protein interactions in yeast [24,27]. Following TAP-isolation, the eluted proteins were separated by SDS–PAGE and visualized by silver staining (Fig. 2A). Affinity purification of Gis2p yielded a large set of co-purifiying proteins, the majority of which were below 40 kDa. We identified the co-purifying peptides and proteins by LC–MS/MS analysis (Supplemental Table 1) [21]. Label-free, semi-quantitative protein abundance factors (PAF) were computed for each identified protein to give a relative measure of the abundance of each protein (Supplemental Table 1 and Supplemental Fig. 1) [28]. Based on this method for comparing LC–MS/MS experiments, the most abundant co-purifying proteins were ribosomal subunits and other translation related proteins (Fig. 2B and Supplemental

Please cite this article in press as: M.A. Sammons et al., *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9, Biochem. Biophys. Res. Commun. (2011), doi:10.1016/j.bbrc.2011.01.086

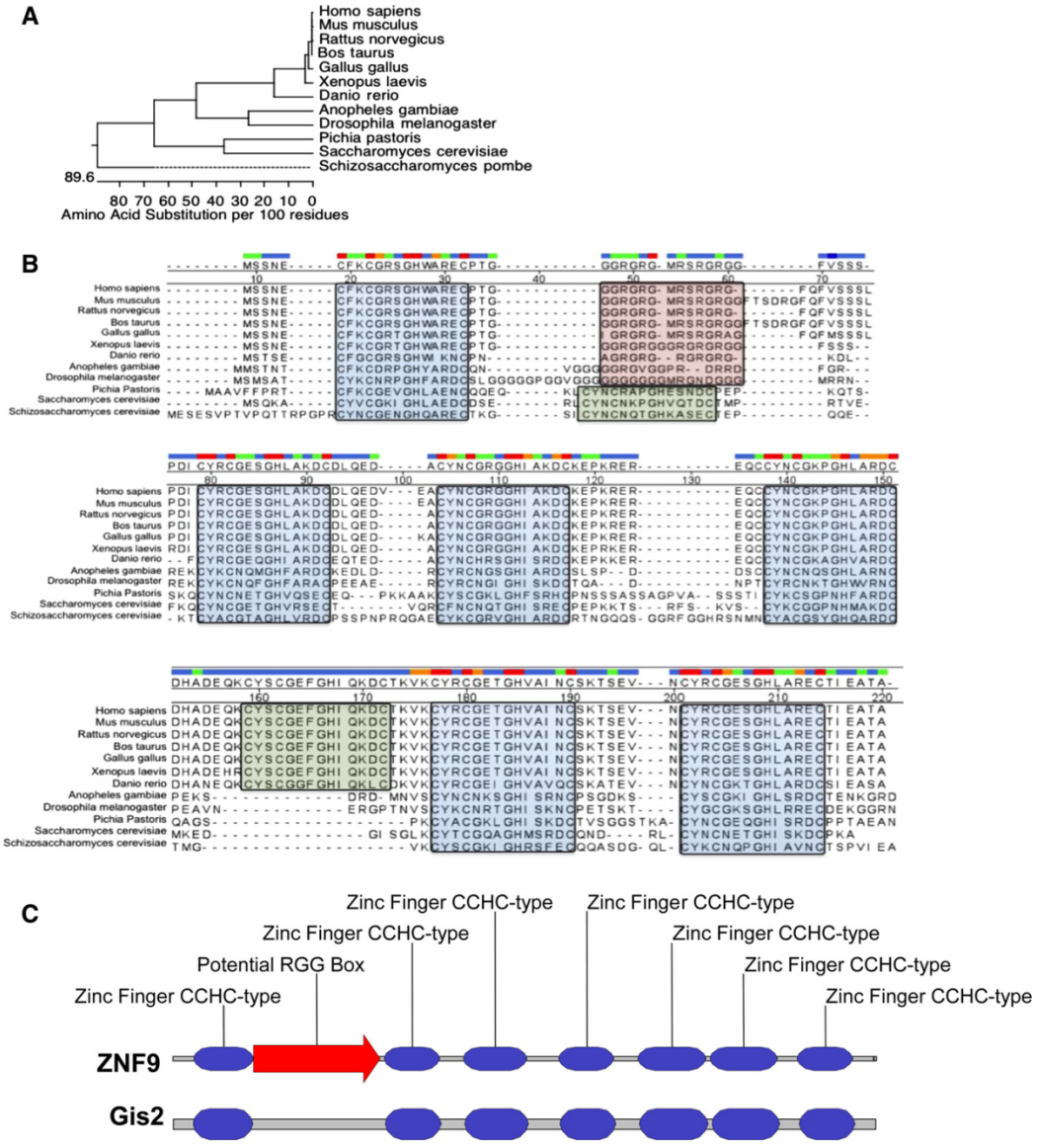



**Fig. 1.** Phylogenetic and sequence analysis of ZNF9/CNBP homologs. (A) RefSeq protein models of ZNF9/CNBP homologs for each organism were analyzed for sequence similarity using the ClustalW algorithm. Divergence and distance are graphed as the number of amino acid differences per 100 residues. (B) Sequence alignment of putative ZNF9 orthologs. Protein sequence alignment of ZNF9/CNBP homologs was performed using the ClustalW algorithm and visualized using MegAlign (DNAStar Lasergene). Blue boxes represent shared CCHC zinc finger motifs found in all protein models. The red box represents the RGG domain, which is only found in multicellular organisms, while the green boxes represent the unaligned CCHC domains differentially located in the unicellular and multicellular groups. (C) Alignment of human ZNF9 and yeast Gis2p functional domains.

Fig. 1). Clustering of data from three replicate Gis2-TAP isolates and four replicate purified ribosomes revealed that Gis2p reproducibly copurifies with ribosomes and other translation factors. In earlier studies, ZNF9 copurifies with a number of ribosomal proteins in HeLa cell lines, and proteomic analysis of immunoprecipitated ZNF9 revealed a large number of copurifiying ribosomal and RNA-binding proteins [8,29]. These data suggest that Gis2p copurifies with components of the eukaryotic ribosome and provide evidence that biochemical interactions are conserved between Gis2p and ZNF9.

Please cite this article in press as: M.A. Sammons et al., *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9, Biochem. Biophys. Res. Commun. (2011), doi:10.1016/j.bbrc.2011.01.086

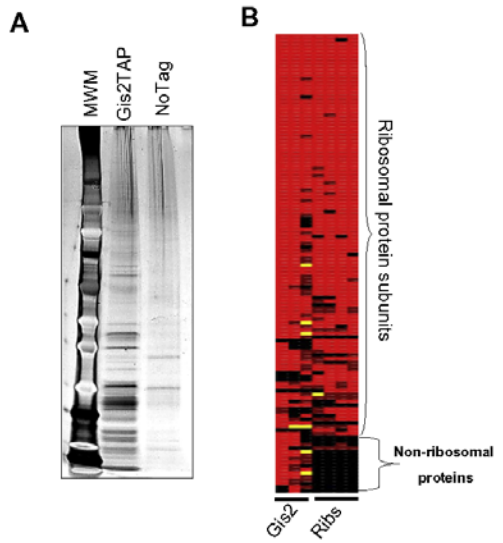



**Fig. 2.** Analysis of proteins isolate by Gis2p-targetted TAP reveals the interactome of Gis2p. (A) Eluted proteins purified by a TAP protocol from a Gis2-TAP expressing strain and an untagged parental strain. 10% of the eluted proteins were analyzed by SDS–PAGE and silver stained (MWM: molecular weight markers). (B) Clustering of proteins identified by tandem mass spectrometry from 3 replicate Gis2p-TAP purifications compared to 4 replicate isolations of ribosomes. On the heat map, protein abundance is represented by color, from black representing no protein identified to bright red meaning high PAF value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 3.2. Polysome analysis reveals Gis2p colocalizes with the ribosome

To independently test whether copurification of Gis2p with ribosomal subunits and the translation machinery is a result of *in vivo* association with the ribosome during translation, we isolated ribosomes by differential centrifugation and assayed for the presence of Gis2p. Initially, lysates from a Gis2-TAP-tagged strain were loaded onto a sucrose cushion and subjected to ultracentrifugation to isolate polyribosomes. As seen in Fig. 3A, the large majority of Gis2p is found in the polysome pellet, which suggests that Gis2p associates with the translating ribosome. We confirmed this result using linear sucrose gradient ultracentrifugation experiments (Fig. 3B). With both centrifugation strategies, a small population of Gis2p was found in the non-ribosomal fractions, suggesting that not all of the cellular Gis2p is actively associated with the ribosome. Human ZNF9 shows the same pattern of localization: predominantly ribosomal, with a small non-ribosome pool [8].

Ribosome salt wash (RSW) experiments are often used to separate core ribosomal subunits from associated translation factors. Previously, we used RSW to determine the strength of human ZNF9's ribosome interaction [8]. We reasoned that Gis2p and ZNF9 should elute from the ribosome at similar salt concentrations if they are biochemical and functional orthologs. As seen previously for human ZNF9, Gis2p remains associated with ribosomes at low salt concentrations (100 mM and 250 mM potassium acetate) (Fig. 3C). With 500 mM salt, Gis2p begins to elute from the ribosome pellet, and at 1 M salt, all of the Gis2p is found in the wash fraction. These results are similar to what we previously observed with human ZNF9 [8]. Gis2p's interaction with the ribosome is similar to that of other translation factors and not quite as strong as that of core ribosomal subunits. The similar ribosomal interactions of Gis2p and ZNF9 provide evidence that they are putative biochemical orthologs and suggest that they function in the essential cellular process of translation.

### 3.3. Gis2p–ribosome interactions are sensitive to RNAse

Next, we investigated the mechanism of Gis2p's interaction with ribosomes. Because of the polysomal localization of Gis2p, we tested whether Gis2p's ribosome association required tethering through RNA by treatment with RNAse A. We treated lysates of Gis2-TAP expressing cells with 50 μg/mL of RNAse. Then, Gis2p was purified and subjected to western blot analysis and mass spectrometry-based proteomics (Fig. 3D). Interestingly, even though more Gis2p was purified from RNAse-treated samples compared to mock-treated controls, less ribosomal protein was identified by mass spectrometry. These results suggest that RNAse A treatment releases Gis2p from the ribosome and makes the purification more efficient. To confirm that RNAse A treatment releases Gis2p from ribosomes, we purified ribosomes were by ultracentrifugation and treated them with RNAse A. As a control, in parallel, we treated ribosomes with EDTA, which disrupts polyribosomes and their interactions with certain RNA binding proteins, including human ZNF9 [8]. Ribosomes were then reisolated through a sucrose cushion and analyzed by western blotting for the presence of Gis2p. As seen in Fig. 3E, treatment of ribosomes with RNAse or EDTA results in a large decrease in Gis2p association, suggesting that Gis2p interacts with actively translating ribosomes and that this interaction requires an interaction with either mRNA or RNAse-sensitive RNA.

### 3.4. Expression of Gis2p activates cap-independent translation of the ODC IRES in HEK293T cells

In analyzing whether Gis2p is an ortholog of ZNF9/CNBP, one of our critical considerations was function. We tested whether a diploid Δ*gis2* yeast strain display slow growth or lethal phenotypes after a number of insults, including a series of translation inhibitors (Supplemental Fig. 2). Growth rate of Δ*gis2* cells was comparable to wild type under all conditions tested, suggesting that Gis2 is non-essential for survival in these conditions. Previously, we showed that ZNF9 functions as an activator of cap-independent translation in mammalian cells, specifically the ODC IRES [7,8]. If Gis2p is a true ortholog of ZNF9, it should facilitate activation of IRES targets. To measure cap-independent translation, we used a previously characterized bicistronic reporter plasmid system (Fig. 4A) [8]. The full-length coding sequence of Gis2p, ZNF9, and an unrelated protein (β-galactosidase) were placed downstream of a V5 epitope tag in a mammalian expression plasmid. Co-expression of the ODC-IRES reporter with V5-ZNF9 in HEK293T cells activated translation of the ODC IRES, whereas co-expression with V5-β-Gal did not (Fig. 4B). Consistent with our hypothesis, expression of V5-Gis2p activated cap-independent translation of the reporter at levels similar to V5-ZNF9 (Fig. 4B). These results show that Gis2p is an activator of IRES-dependent ODC translation and provide further evidence that Gis2p and ZNF9 are functional and biochemical orthologs.

### 3.5. A conserved C-terminal region of ZNF9 is required for full activation of cap-independent translation

We hypothesized that regions of Gis2p critical for its cap-independent translation activity are in evolutionarily conserved regions or domains shared with ZNF9. Gis2p contains 7 sequential CCHC-type zinc finger motifs and no other conserved functional domains. ZNF9 additionally contains an RGG box motif between

Please cite this article in press as: M.A. Sammons et al., *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9, Biochem. Biophys. Res. Commun. (2011), doi:10.1016/j.bbrc.2011.01.086

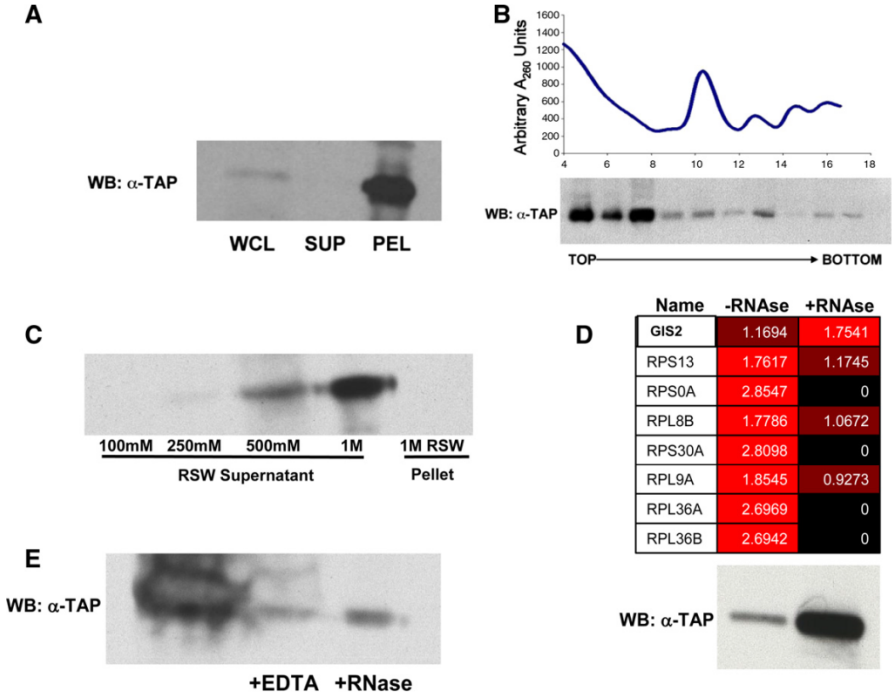



**Fig. 3.** *S. cerevisiae* Gis2p interacts with the ribosome. (A) Western blot analysis for the presence of Gis2p in whole cell lysates (WCL), supernatant (SUP), or pellet (PEL) fractions after isolation of polysomes from Gis2-TAP-tagged yeast cells. (B) Polysome analysis of the Gis2-TAP strain was measured using on-line $UV_{260}$ absorbance. 10% of each fraction was analyzed by western blotting for the presence of Gis2p. (C) Western analysis for Gis2p in ribosome salt washes of increasing concentrations of potassium acetate. (D) Changing abundance of a subset of proteins isolated by purification of Gis2-TAP in the presence or absence of RNAse. (E) Western analysis for Gis2-TAP in solubilized polysome pellets treated with EDTA or RNAse.

the first and second zinc fingers. To test the regions of ZNF9 that are required for cap-independent translation activity, we created a series of N and C-terminal deletion mutants (Fig. 4C). Western analysis of the various ZNF9 constructs in HEK293T cells showed equivalent expression of the Wt and deletion constructs (Fig. 4D). Each mutant was co-expressed with the ODC IRES reporter in HEK293T cells, and cap-independent translation activity was measured (Fig. 4E). Deletion of either the first CCHC-zinc finger or the first zinc finger in combination with the RGG box had no effect on the ability of ZNF9 to activate cap-independent translation. The Δ5 mutant, lacking the fifth, sixth, and seventh zinc fingers, fails to fully activate cap-independent translation, whereas the Δ6 mutant, lacking only the sixth and seventh zinc fingers, shows no difference in activity compared to the wild-type ZNF9. These data demonstrate that the RGG box of ZNF9, which is not found in Gis2p, is not important for cap-independent translation activity, but that an intact fifth CCHC-zinc finger is necessary for full activation of translation.

## 4. Discussion

### 4.1. Gis2p has evolutionarily conserved functions in translation

Because proteins are the critical catalysts of all cellular functions, eukaryotic cells have sophisticated and diverse mechanisms to temporally and spatially control protein synthesis. This work provides evidence that the *S. cerevisiae* protein Gis2p is part of an evolutionarily conserved mechanism to control translation of specific mRNA molecules that contain internal ribosome entry sites. Multiple studies have demonstrated cap-independent translation in *S. cerevisiae* of both viral and endogenous cellular mRNA molecules [30–33]. IRES-mediated translation of viral mRNAs is thought to occur through conserved secondary structure inherent to the IRES itself, but IRES-mediated translation of cellular RNAs is believed to require sequence specific cofactors, or ITAFs. It remains to be seen whether Gis2p functions as an ITAF for any of the known yeast IRES-containing mRNAs. The conserved function of Gis2p in mammalian IRES-mediated translation suggests that Gis2p likely acts as an ITAF in yeast. No mRNA-specific ITAFs have been identified to date in yeast [30–32,34]. Further investigation into what, if any, yeast cellular IRESs are regulated by ITAFs, and Gis2p in particular, will be needed to fully understand the physiological role of cap-independent translation in yeast and other eukaryotic organisms.

### 4.2. Gis2p and ZNF9 share common biochemical interactions

As demonstrated here and previously [7,8], Gis2p and ZNF9 associate with the eukaryotic ribosome and act as regulators of cap-independent translation. ZNF9 interacts with other known ITAFs, such as PCBP2 and the La protein [7,12,14,35], which may have vital roles in regulating the cap-independent translation activity of ZNF9. Interestingly, the *S. cerevisiae* protein Sro9p is a putative ortholog of the mammalian La protein and was identified

Please cite this article in press as: M.A. Sammons et al., *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9, Biochem. Biophys. Res. Commun. (2011), doi:10.1016/j.bbrc.2011.01.086

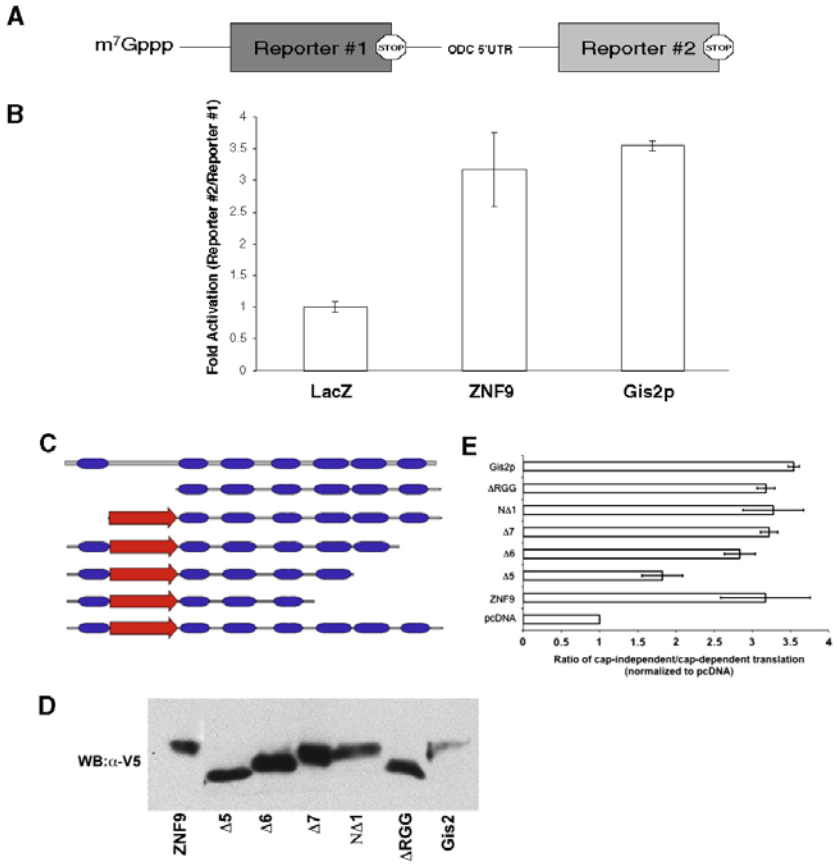



**Fig. 4.** Expression of Gis2p and ZNF9 deletion mutants in HEK293T cells activates IRES-dependent translation. (A) Schematic representation of the bicistronic ODC-IRES reporter. (B) Cap-independent translation in HEK293T cells expressing the bicistronic ODC-IRES reporter and either LacZ, ZNF9, or Gis2p. Cells were analyzed for luciferase activity, and cap-independent translation is reported as a ratio of cap-independent translation (reporter #2, firefly luciferase) to cap-dependent translation (reporter #1, *Renilla* luciferase). Three independent experiments were averaged. The error bars show the standard deviations. (C) Schematic representation of Gis2p and various ZNF9 deletion mutants. (D) Western blot analysis using anti-V5 monoclonal antibody against V5-tagged ZNF9 constructs and Gis2p protein expression in HEK293T cells (the lane corresponding to the wild-type ZNF9 construct was cropped from another gel for clarity). (E) Cap-independent translation in HEK293T cells expressing the ODC-IRES reporter and the Gis2p and ZNF9 plasmid constructs. The plasmid constructs shown in C. are aligned with the translation assay results, respectively. Cells were analyzed for cap-independent activity as described above. Three independent experiments were averaged. The error bars represent the standard deviations.

as a potential Gis2p interactor in our studies (Supplemental Fig. 2). The Michnick group recently showed a putative interaction between Sro9p and Gis2p by genome-wide protein-fragment complementation assay [36]. Sro9p was initially identified as a multi copy suppressor of an RNA export defect and was shown to associate with polysomes [37,38]. Sro9p associates with nascent transcripts and forms part of an mRNP complex that is exported to the cytoplasm, where Sro9p then acts as a regulator of translation [39].

### *4.3. Using S. cerevisiae as a model system to study ZNF9 function and the regulation of cap-independent translation*

The yeast model system offers intriguing opportunities to study Gis2p and Sro9p and their potentially conserved cellular roles in translation. The eukaryotic translation machinery is extremely well-conserved from yeast to man. The rapid growth rate and the ease of identifying mutants make yeast a very tractable model system in which to study the essential cellular process of protein synthesis. While the role that ZNF9 plays in the progression of DM2 is best studied in the context of mammalian cells, the role of ZNF9 in cap-independent translation can be addressed using a variety of approaches in yeast to identify novel regulators and functions of the ZNF9 homolog, Gis2p. As previously discussed, the conservation of ITAF activity between ZNF9 and Gis2p suggests that the molecular and biochemical interactions and activities of Gis2p in yeast are directly relevant in understanding ZNF9 function.

Although previous work suggested that ZNF9 plays a role in cell proliferation, Δ*gis2* deletion mutants grew at a rate similar to the wild type rate under all conditions tested, suggesting that ZNF9's role in cell proliferation is not shared by Gis2p. One interesting possibility is that the RGG box of ZNF9, which is not found in Gis2p, functions in the control of cell proliferation and growth [6]. Further studies are needed to understand what role the different conserved

Please cite this article in press as: M.A. Sammons et al., *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9, Biochem. Biophys. Res. Commun. (2011), doi:10.1016/j.bbrc.2011.01.086

and unique functional domains play in the multiple roles of ZNF9 and Gis2p. Collectively, the data presented here show that Gis2p is a functional ortholog of the mammalian ZNF9 protein and suggest that information gleaned from the study of Gis2p in yeast is relevant to the understanding of ZNF9.

## Acknowledgments

We acknowledge Elizabeth M. Link for helpful comments and suggestions. M.A.S. was supported by NIH grant R21 AR055231. P.S. was supported by R01 GM64779. A.J.L. was supported by NIH grants R21 AR055231 and R01 GM64779.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2011.01.086.

## References

[1] C.L. Liquori, K. Ricker, M.L. Moseley, J.F. Jacobsen, W. Kress, S.L. Naylor, J.W. Day, L.P. Ranum, Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9, Science 293 (2001) 864–867.
[2] L. Pellizzoni, F. Lotti, B. Maras, P. Pierandrei-Amaldi, Cellular nucleic acid binding protein binds a conserved region of the 5′ UTR of *Xenopus laevis* ribosomal protein mRNAs, J. Mol. Biol. 267 (1997) 264–275.
[3] W. Chen, Y. Liang, W. Deng, K. Shimizu, A.M. Ashique, E. Li, Y.P. Li, The zinc-finger protein CNBP is required for forebrain formation in the mouse, Development 130 (2003) 1367–1379.
[4] Y. Abe, W. Chen, W. Huang, M. Nishino, Y.P. Li, CNBP regulates forebrain formation at organogenesis stage in chick embryos, Dev. Biol. 295 (2006) 116–127.
[5] A.M. Weiner, M.L. Allende, T.S. Becker, N.B. Calcaterra, CNBP mediates neural crest cell expansion by controlling cell proliferation and cell survival during rostral head development, J. Cell Biochem. 102 (2007) 1553–1570.
[6] P. Armas, T.H. Aguero, M. Borgognone, M.J. Aybar, N.B. Calcaterra, C.N.B.P. Dissecting, A zinc-finger protein required for neural crest development, in its structural and functional domains, J. Mol. Biol. 382 (2008) 1043–1056.
[7] V.R. Gerbasi, A.J. Link, The myotonic dystrophy type 2 protein ZNF9 is part of an ITAF complex that promotes cap-independent translation, Mol. Cell Proteomics 6 (2007) 1049–1058.
[8] M.A. Sammons, A.K. Antons, M. Bendjennat, B. Udd, R. Krahe, A.J. Link, ZNF9 activation of IRES-mediated translation of the human ODC mRNA is decreased in myotonic dystrophy type 2, PLoS One 5 (2010) e9301.
[9] S.D. Baird, M. Turcotte, R.G. Korneluk, M. Holcik, Searching for IRES, RNA 12 (2006) 1755–1785.
[10] K.D. Fitzgerald, B.L. Semler, Bridging IRES elements in mRNAs to the eukaryotic translation apparatus, Biochim. Biophys. Acta 1789 (2009) 518–528.
[11] A. Pacheco, E. Martinez-Salas, Insights into the biology of IRES elements through riboproteomic approaches, J. Biomed. Biotechnol. 2010 (2010) 458927.
[12] L. Pellizzoni, F. Lotti, S.A. Rutjes, P. Pierandrei-Amaldi, Involvement of the *Xenopus laevis* Ro60 autoantigen in the alternative interaction of La and CNBP proteins with the 5′UTR of L4 ribosomal protein mRNA, J. Mol. Biol. 281 (1998) 593–608.
[13] C. Crosio, P.P. Boyl, F. Loreni, P. Pierandrei-Amaldi, F. Amaldi, La protein has a positive effect on the translation of TOP mRNAs in vivo, Nucleic Acids Res. 28 (2000) 2927–2934.
[14] S. Schlatter, M. Fussenegger, Novel CNBP- and La-based translation control systems for mammalian cells, Biotechnol. Bioeng. 81 (2003) 1–12.
[15] C. Huichalaf, B. Schoser, C. Schneider-Gold, B. Jin, P. Sarkar, L. Timchenko, Reduction of the rate of protein translation in patients with myotonic dystrophy 2, J. Neurosci. 29 (2009) 9042–9049.
[16] D. Balciunas, H. Ronne, Yeast genes GIS1-4: multicopy suppressors of the Gal-phenotype of snf1 mig1 srb8/10/11 cells, Mol. Gen. Genet. 262 (1999) 589–599.
[17] K. Shimizu, W. Chen, A.M. Ashique, R. Moroi, Y.P. Li, Molecular cloning, developmental expression, promoter analysis and functional characterization of the mouse CNBP gene, Gene 307 (2003) 51–62.
[18] D.C. Amberg, D.J. Burke, J.N. Strathern, Methods in Yeast Genetics 2005 Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2005.
[19] E.A. Winzeler, D.D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J.D. Boeke, H. Bussey, et al., Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis, Science 285 (1999) 901–906.
[20] S. Ghaemmaghami, W.K. Huh, K. Bower, R.W. Howson, A. Belle, N. Dephoure, E.K. O'Shea, J.S. Weissman, Global analysis of protein expression in yeast, Nature 425 (2003) 737–741.
[21] A.J. Link, T.C. Fleischer, C.M. Weaver, V.R. Gerbasi, J.L. Jennings, Purifying protein complexes for mass spectrometry: applications to protein translation, Methods 35 (2005) 274–290.
[22] J.K. Eng, A.L. McCormack, I.J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences, J. Am. Soc. Mass Spectrom. 5 (1994) 976–989.
[23] K.J. McAfee, D.T. Duncan, M. Assink, A.J. Link, Analyzing proteomes and protein function using graphical comparative analysis of tandem mass spectrometry results, Mol. Cell Proteomics 5 (2006) 1497–1513.
[24] T.C. Fleischer, C.M. Weaver, K.J. McAfee, J.L. Jennings, A.J. Link, Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes, Genes Dev. 20 (2006) 1294–1307.
[25] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[26] T.B. Rajavashisth, A.K. Taylor, A. Andalibi, K.L. Svenson, A.J. Lusis, Identification of a zinc finger protein that binds to the sterol regulatory element, Science 245 (1989) 640–643.
[27] N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, et al., Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*, Nature 440 (2006) 637–643.
[28] D.W. Powell, C.M. Weaver, J.L. Jennings, K.J. McAfee, Y. He, P.A. Weil, A.J. Link, Cluster analysis of mass spectrometry data reveals a novel component of SAGA, Mol. Cell Biol. 24 (2004) 7249–7259.
[29] R.M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M.D. Robinson, L. O'Connor, M. Li, et al., Large-scale mapping of human protein–protein interactions by mass spectrometry, Mol. Syst. Biol. 3 (2007) 89.
[30] S.R. Thompson, K.D. Gulyas, P. Sarnow, Internal initiation in *Saccharomyces cerevisiae* mediated by an initiator tRNA/eIF2-independent internal ribosome entry site element, Proc. Natl. Acad. Sci. USA 98 (2001) 12972–12977.
[31] A.A. Komar, T. Lesnik, C. Cullin, W.C. Merrick, H. Trachsel, M. Altmann, Internal initiation drives the synthesis of Ure2 protein lacking the prion domain and affects [URE3] propagation in yeast cells, EMBO J. 22 (2003) 1199–1209.
[32] A.A. Komar, M. Hatzoglou, Internal ribosome entry sites in cellular mRNAs: mystery of their existence, J. Biol. Chem. 280 (2005) 23425–23428.
[33] W.V. Gilbert, K. Zhou, T.K. Butler, J.A. Doudna, Cap-independent translation is required for starvation-induced differentiation in yeast, Science 317 (2007) 1224–1227.
[34] D.M. Landry, M.I. Hertz, S.R. Thompson, RPS25 is essential for translation initiation by the Dicistroviridae and hepatitis C viral IRESs, Genes Dev. 23 (2009) 2753–2764.
[35] B. Cardinali, C. Carissimi, P. Gravina, P. Pierandrei-Amaldi, La protein is associated with terminal oligopyrimidine mRNAs in actively translating polysomes, J. Biol. Chem. 278 (2003) 35145–35151.
[36] K. Tarassov, V. Messier, C.R. Landry, S. Radinovic, M.M. Serna Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey, S.W. Michnick, An in vivo map of the yeast protein interactome, Science 320 (2008) 1465–1470.
[37] M. Kagami, A. Toh-e, Y. Matsui, SRO9, a multicopy suppressor of the bud growth defect in the *Saccharomyces cerevisiae* rho3-deficient cells, shows strong genetic interactions with tropomyosin genes, suggesting its role in organization of the actin cytoskeleton, Genetics 147 (1997) 1003–1016.
[38] S.G. Sobel, S.L. Wolin, Two yeast La motif-containing proteins are RNA-binding proteins that associate with polyribosomes, Mol. Biol. Cell 10 (1999) 3849–3862.
[39] S. Rother, C. Burkert, K.M. Brunger, A. Mayer, A. Kieser, K. Strasser, Nucleocytoplasmic shuttling of the La motif-containing protein Sro9 might link its nuclear and cytoplasmic functions, RNA 16 (2010) 1393–1401.

Please cite this article in press as: M.A. Sammons et al., *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9, Biochem. Biophys. Res. Commun. (2011), doi:10.1016/j.bbrc.2011.01.086

# Appendix AC – Manuscript – 8: Analyzing the Cryptome: Uncovering Secret Sequences

*The AAPS Journal, Vol. 13, No. 2, June 2011 (© 2011)*
DOI: 10.1208/s12248-011-9252-2

*Mini-Review*

*Theme: Fishing for the Hidden Proteome in Health and Disease: Focus on Drug Abuse*
*Guest Editors: Rao S. Rapaka, Lloyd D. Fricker, and Jonathan V. Sweedler*

# Analyzing the Cryptome: Uncovering Secret Sequences

**Parimal Samir[1] and Andrew J. Link[1,2,3]**

*Received 29 August 2010; accepted 23 December 2010; published online 16 February 2011*

***Abstract.*** The mammalian cryptome consists of bioactive peptides generated by the proteolysis of precursor proteins. It is speculated that the cryptide repertoire increases the complexity of the proteome by an order of magnitude. Cryptides have been found to function in a wide range of processes including neuronal signaling, antigen presentation, and the inflammatory response. Due to their potential as therapeutic agents, there has been an increasing interest in studying cryptides. In this review, we discuss different approaches for discovering these hidden peptides and how proteomic tools can be utilized to aid in their identification and characterization.

**KEY WORDS:** bioactive peptides; cryptides; cryptome; cryptomics; mass spectrometry-based proteomics.

## INTRODUCTION: WHAT IS CRYPTOMICS AND WHY IS IT IMPORTANT?

A typical mammalian proteome has tens of thousands of unique proteins. It is staggering to imagine understanding the detailed function of each one. Yet a simple count of the proteins actually underestimates the true complexity of a proteome. Proteins have numerous isoforms. They are modified with a variety of transcriptional and posttranslational modifications. Each variant has a potentially important biological function (1). An added layer of complexity is introduced by proteolytic enzymes that cleave precursor proteins to generate bioactive peptides. These peptides have been termed cryptides, and their study *en mass* is known as cryptomics (2–4). It is important to note that all peptides generated from proteins are not considered cryptides; a cryptide has a biological activity distinct from its precursor. For example, upon activation of the Notch receptor protein, its cytoplasmic domain is proteolytically cleaved and migrates to the nucleus to regulate the Notch target genes. The cytosolic fragment of Notch is not considered a cryptide since it carries out the main function of the Notch protein (5,6). In contrast, for example, a cryptide is generated when cytochrome c oxidase subunit VIII is cleaved to create mitocryptide-1. The peptide has a distinct function as an activator of neutrophils (Fig. 1a) (7). Another example of a cryptide is parstatin, the N-terminal extracellular domain fragment of proteinase activated receptor 1 (PAR1), which is formed upon cleavage of PAR1 by thrombin. The N-terminal domain has an unexpected function as an inhibitor of angiogenesis and thus has the capability to antagonize the pro-angiogenesis function of its precursor PAR1 (Fig. 1b) (8).

The first proof that a bioactive peptide can be the proteolytic product of a larger precursor was provided by the study of proopiomelanocortin, which gives rise to multiple cryptides, including ACTH, β-endorphin, lipotrophins, and melanocyte-stimulating hormones (9–13). Since then a large number of cryptides and their corresponding precursors have been identified along with various proteases that are involved. Recently discovered cryptides are involved in the regulation of a diverse range of cellular processes including neuronal signaling, the inflammatory response, the adaptive immune response, and angiogenesis (Table 1). They are derived from many well-studied proteins, including hemoglobin, cytochromes, laminins, and collagen (2–4,18,19). However, the occurrence and function of cryptides cannot currently be accurately predicted. As such, cryptides are a hidden aspect of the proteome with important yet unknown biological functions. As an evidence of the biodiversity of cryptides, a glycine/leucine-rich bioactive peptide with antimicrobial activity, leptoglycin, has been discovered in the South American frog, although the precursor is still to be identified (14).

Cryptides have been divided into three types by Dominic Autelitano and co-workers (2). A type I cryptide is a bioactive peptide detected *in vivo* with a function entirely different from that of the precursor. A type II cryptide is a peptide found *in vivo* with activity related but not necessarily identical to that of its precursor. Finally, a type III cryptide is a bioactive peptide produced *in vitro* by proteolytic digestion of proteins that may or may not exist *in vivo* (2). This classification scheme organizes the rapidly evolving area of

[1] Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, Tennessee 37232-2363, USA.
[2] Departments of Microbiology and Immunology, Vanderbilt University School of Medicine, 1161 21st Ave South, Nashville, Tennessee 37232-2363, USA.
[3] To whom correspondence should be addressed. (e-mail: andrew.link@vanderbilt.edu)

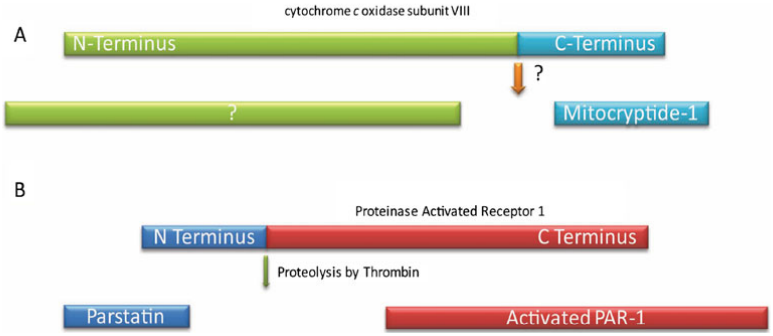



**Fig. 1.** **a** Mitocryptide-1 is the 23-amino acid C-terminal fragment of cytochrome *c* oxidase subunit VIII originally identified from porcine heart. The protease that cleaves the precursor protein has not been identified yet. Mitocryptide-1 has been found to activate neutrophils. The fate of N-terminal fragment is not known. **b** Parstatin is the N-terminal fragment of proteinase activated receptor 1 (*PAR1*). Parstatin is generated when thrombin cleaves PAR1 for its activation. The cryptide has anti-angiogenic activity

bioactive peptide research. In this review, we discuss recent progress made since the concept of cryptome was first proposed in 2006. We will also discuss different approaches for the discovery of novel cryptides and how mass spectrometry-based proteomics has been and can be utilized in cryptomic research.

Thus far, cryptides have been identified and characterized individually. They are varied in their origins and functions. Type I cryptides are derived from plasma proteins, extracellular matrix proteins, cytosolic, and mitochondrial proteins, among others. For example, plasminogen, a clotting factor, is cleaved to yield angiostatin, a 38-kDa cryptide that is a potent inhibitor of angiogenesis (20). Collagen VIII is cleaved by metalloproteases to generate endostatin, a 20-kDa cryptide with tumor repression activity.(21). Cleavage of hemoglobin results in a variety of cryptides, including hemopressins and hemorphins. Some of these are involved in neuronal signaling (reviewed in (18,19)) and others have antimicrobial activity (22,23). Type II and type III cryptides have also been found from a range of sources and have equally complex biological functions. An example of a type II cryptide is NPNA. The rat neuropeptide FF is cleaved to form a novel neuropeptide NPNA. NPNA influences opioid receptor signaling by reducing the mRNA expression of G-protein α subunits associated with opioid receptors (17,24). As an example of type III cryptides, cytochrome c can be cleaved to generate cell penetrating peptides that have potent apoptogenic activity (25). The apoptogenicity of one of the peptides, Cyt $c^{77\ 101}$, was increased by constructing a chimeric synthetic cryptide with a nonapeptide N-terminal extension derived from the C-terminal region of nucleoporin 153. The nucleoporin 153 is a component of nuclear pore complex containing FG repeat at its C terminus. The modified cryptide has apoptogenic activity at sub-micromolar levels, the range of concentration readily achievable for therapeutics.

Although cryptomics is in its infancy, cryptides are already known to originate from a variety of precursors and to possess a vast range of bioactivities. It is not difficult to imagine many more cryptides being generated from unexpected candidate proteins. There is an increasing interest in developing new methodologies to identify and characterize cryptides because of their diverse roles and their potential for therapeutic use (2). The identification of novel cryptides combined with quantitative studies to measure the concentration of individual or sets of cryptides in cells and tissues is expected to answer fundamental questions about their function and regulation.

## DISCOVERY OF CRYPTIDES

Varied approaches have been employed to discover novel cryptides. One method involves using a diagnostic assay to screen peptides for a particular biological activity. For example, Daniel Pimenta and co-workers isolated low molecular weight components of dog pancreas and analyzed fractions for bradykinin potentiating activity (26). In this way, they identified cryptides with properties similar to hemorphins and hemorphin-like cryptides, which are derived from

**Table 1.** Examples of Recently Discovered Cryptides with Their Precursors

| Cryptide | Number of amino acids | Precursor | Cellular role |
|---|---|---|---|
| Leptoglycin | 22 | Unknown | Antimicrobial activity (14) |
| Parstatin | 41 | PAR1 | Inhibition of angiogenesis (8) |
| Prolyl-hydroxyproline | 2 | Collagen | Inhibition of chondrocyte differentiation (15) |
| Mytocryptide-1 | 23 | Cytochrome c oxidase subunit VIII | Activation of neutrophils (7) |
| Mytocryptide-2 | 15 | Cytochrome b | Activation of neutrophils (16) |
| NPNA | 15 | Rat neuropeptide FF precursor | Opioid signaling (17) |

*PAR1* proteinase activated receptor 1

hemoglobin (26). A number of other screening approaches, including screening proteolytic digests of purified proteins and screening synthetic random peptide pools, have been successfully applied to identify novel cryptides (3).

Computational biology techniques can be used to predict possible cryptides. For example, Hidehito Mukai and co-workers used the previously identified neutrophil-activating factors mastoparan and mitocryptide-1 as the basis for identifying novel neutrophil-activating factors (4). Mitocryptide-1 is derived from cytochrome c oxidase subunit VIII. Their first step was to generate a list of 15–36 residue peptide fragments predicted to result from the action of various mitochondrial peptidases and cellular proteases on the 441 human mitochondrial proteins (4). They compared the physicochemical properties and three-dimensional structures of the peptides to mastoparan and mitocryptide-1 to generate a list of putative neutrophil-activating peptides. Finally, they synthesized the selected peptides and assayed their ability to activate neutrophils. They identified eight novel peptides that activate neutrophils, including several peptide fragments from cytochrome c (4). To identify cryptides involved in cell adhesion, Yoshihiko Yamada and co-workers synthesized a series of predicted proteolytic fragments of laminin and assayed their biological activity (27–30). Their systematic screen yielded an extensive list of different laminin proteolytic fragments that may affect cell adhesion (27–30).

## MASS SPECTROMETRY-BASED PROTEOMIC APPROACHES FOR THE DISCOVERY OF CRYPTIDES

Mass spectrometry-based approaches are powerful and comprehensive tools for analyzing the cryptome. These approaches have been extensively used to qualitatively and quantitatively characterize predicted proteomes in multitudes of organisms, tissues, and cell types. Mass spectrometry-based cryptomics can be conceptually utilized in two ways: (a) to search for cryptides in homogenized tissue and (b) to identify and determine the tissue distribution of cryptides *in situ* (Figs. 2 and 3). In the following sections, we will discuss the mass spectrometry-based technologies that can be exploited to identify novel cryptides from biological samples, borrowing mainly from the technologies utilized for peptidome profiling. These methods are expected to reveal a variety of peptides present in a sample, which can be tested subsequently for their bioactivity. Moreover, quantitative proteomic tools now allow the simultaneous quantitation of multiple cryptides with high sensitivity and accuracy in a single experiment.

The development of efficient peptide fractionation using liquid chromatography coupled to electrospray ionization has led to a tremendous capability for peptide identification and quantification. Multi-dimensional protein identification technology (MuDPIT) has been utilized in a number of large-scale proteome profiling studies (32–34), and variations have been specifically used to profile endogenous small peptides. For example, MuDPIT-based strategies have been utilized to profile the peptidome of urine and to compare the self-antigen peptidomes of plasma and lymphatic fluid (35,36). A number of studies have analyzed the blood peptidome, many with the goal of discovering cancer biomarkers (37). These approaches have identified many low molecular weight

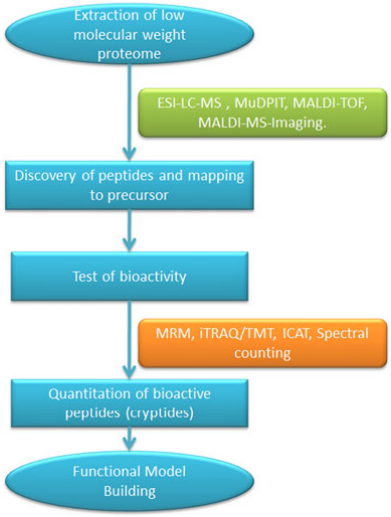



**Fig. 2.** Mass spectrometry-based cryptide discovery and quantitation work flow: Cryptomics work flow for the high throughput discovery of cryptides would start with fractionation of tissue to enrich for low molecular weight proteome. A variety of mass spectrometry approaches can be used to identify the enriched cryptides. In most cases, the identified peptides can be mapped to the precursor proteins. Putative cryptides are tested for their bioactivities using either biochemical or cell-based assays. After the validation of bioactivity, the cryptides can be quantitated using a variety of approaches. The advantage of using mass spectrometry-based proteomic tools for quantitation is that large numbers of cryptides can be accurately quantitated in a single experiment. This information can be utilized for building higher order models of the functions of the cryptides. *LC* liquid chromatography, *MS* mass spectrometry, *MALDI* matrix-assisted laser desorption ionization, *ESI* electrospray ionization, *iTRAQ* isobaric tag for absolute and relative quantitation, *TMT* tandem mass tag, *ICAT* isotope coded affinity tags, *MuDPIT* multi-dimensional protein identification technology, *TOF* time of flight, *MRM* multiple reaction monitoring

components of the blood proteome, many of which are expected to be cryptides (37).

Electrospray ionization-Fourier transform mass spectrometry (ESI-FTMS) has been used to identify proteolytic cleavage products of the plasma proteome. Yufeng Shen and co-workers first depleted the 12 most abundant plasma proteins and used size-exclusion chromatography to enrich for peptides with a molecular weight less than 20 kDa (38). Then, using ultra-high-pressure liquid chromatography (UHPLC) ESI-FTMS on a LTQ-Orbitrap mass spectrometer, they identified more than 200 peptides from 29 precursors (38). Using this UHPLC-ESI-FTMS strategy, they also performed a peptidomic profiling of yeast whole cell lysate and have identified about 1,100 peptides from approximately 200 precursor proteins (39).

Similar strategies have also been used to identify novel candidate cryptides from tissue homogenates. Per Andrén and co-workers used nanoLC electrospray ionization quadrupole time-of-flight mass spectrometry to identify peptides from brain tissue (40). Their innovative enrichment for

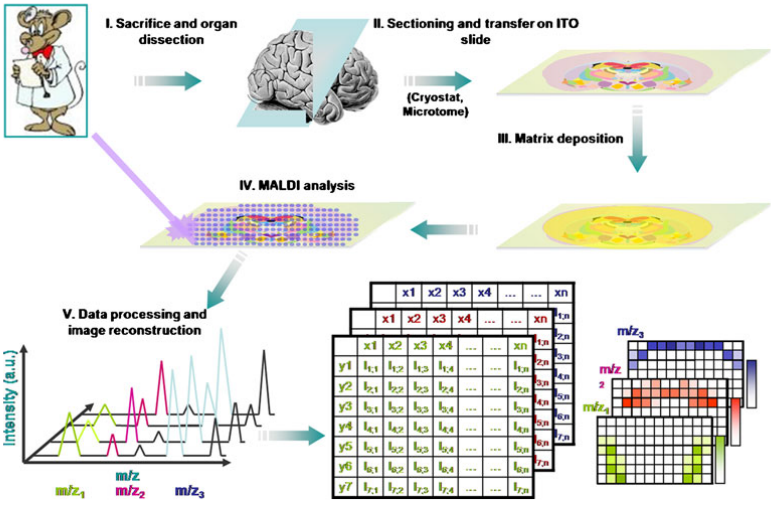



**Fig. 3.** Overview of MALDI-MS imaging for tissue-based cryptomics. In MALDI-MS imaging, 10–20-μm-thick frozen tissue sections are cut using a microtome and transferred to a conductive surface. After matrix is deposited onto the section, MALDI ionization is carried out in the defined regions of the tissue section and spectral data acquired. The mass spectrometry data are computationally processed to generate a pseudoimage, showing abundances of individual peptides and proteins. *a.u.* arbitrary units, *ITO* iridium tin oxide (reprinted with permission from (31))

possible cryptides employed a microfilter device to remove proteins larger than 10 kDa from the mixture. They analyzed the filtered sample using ESI-MS to measure the masses of the peptides. They were able to detect approximately 1,500 endogenous brain peptides. Using tandem mass spectrometry, they were able to identify the sequences of 10% of the detected peptides, including novel peptides (40). In another study, Hanfa Zou and co-workers used ultrafiltration followed by size-exclusion chromatography to fractionate the low molecular weight proteome of liver (41). They directly analyzed the fractions with peptides smaller than 3 kDa and digested the fractions with larger molecules with trypsin before analysis. Using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF-MS) and LC-ESI-MS, they were able to identify more than 1,000 peptides from 400 precursor proteins (41). A similar strategy has been used by William Hancock and co-workers to generate a serum peptidome profile (42).

Peptidome profiling from tissue homogenates is a highly sensitive and fast approach for identifying and quantitating a large number of molecules. But often, one is interested in the tissue-specific distribution of peptides. Moreover, the levels of particular peptides may not be constant throughout a given tissue. The resulting dilution in a tissue homogenate makes it harder to identify them. Therefore, peptidome profiling of specific tissue regions or cells not only helps in understanding the tissue-specific distribution but also allows the identification of additional peptides that may be at elevated levels only in the small regions of the tissue. One technique that has been extensively utilized for visualizing tissue-specific distribution of peptides and proteins is MALDI-MS imaging (Fig. 3) (43). The technology and applications of MALDI-MS imaging have been comprehensively reviewed in (31,44). Recently, a number of studies have used MALDI-TOF-MS to analyze single cells (45,46). These studies demonstrate the potential for utilizing the mass spectrometry-based approaches for the discovery of novel cryptides.

## MASS SPECTROMETRY-BASED PROTEOMIC APPROACHES FOR QUANTITATION OF CRYPTIDES

Proteomic tools have grown beyond their application for discovery of novel targets and can be used to generate quantitative profiles of cryptides. A number of techniques have been developed in the past several years that can determine relative levels of peptides and proteins using mass spectrometry (47). For quantitation, the power of mass spectrometry-based proteomics lies in its ability to detect a wide dynamic range of peptide concentrations. These proteomic approaches include label-free methods, metabolic labeling with stable isotopes, and post-lysis labeling of peptides with isotopic tags (48). Because levels of cryptides are likely to change in response to external stimuli or signaling (2,3), quantitative proteomic tools will be important for characterizing the biological roles of cryptides.

Selected reaction monitoring/multiple reaction monitoring (SRM/MRM) is a targeted mass spectrometry approach to identify and/or quantify peptide(s) and can be utilized for quantitating cryptide(s) that is known *a priori* (49). This is a very powerful technique, with the highest dynamic range and sensitivity of all mass spectrometry-based proteomic tools (50). In addition, SRM/MRM can be used for absolute quantitation of cryptides. In this method, a cryptide can be chosen based on prior mass spectrometry data or *in silico* proteolytic digestion of proteins. Fragmentation spectra of the target cryptide can be either generated or extracted from

spectral libraries. An internal standard peptide (corresponding to the sequence identical to the cryptide of interest) is synthesized with a stable isotope label, and a known amount is spiked into the sample, a strategy known as stable isotope dilution (51). Reporter fragment ion peaks, called transitions, are selected, and peak areas for the heavy and light cryptides are compared to extract quantitative information. Since the amount of the internal standard is known, MRM allows absolute quantitation of the cryptide in a sample (49,52). SRM/MRM is expected to have a great potential for measuring the changing abundance of cryptides in cells and tissues and also for detecting computationally predicted cryptides.

Isobaric tag for relative and absolute quantitation (iTRAQ) and tandem mass tag (TMT) are approaches that use isobaric mass tags that react with the amino group at the N terminus and the epsilon amino group of lysine residues (53,54). Both iTRAQ and TMT reagents have three functional groups, an amine reactive group that reacts with primary amines, a mass balancer region that allows different tags to have the same mass, and a reporter region whose mass varies between different tags. In this approach, two or more samples are allowed to react with the tag, mixed together, processed, and analyzed using tandem mass spectrometry. Identical cryptides with isobaric mass tags from different samples are expected to migrate together during chromatography, and since they are isobaric, they will be selected simultaneously for fragmentation inside the mass spectrometer. Upon fragmentation, the reporter ions will be released from the peptide in the low mass region of the spectrum. Quantitative information can be gleaned by comparing the intensities of the different reporter ions in the same spectra. The spectrum also provides sequence information for the identification of the cryptide. iTRAQ has been multiplexed up to eight channels, while TMT has been multiplexed up to six channels (55,56). Isobaric mass tagging methods are expected to be useful for quantifying the low molecular weight cryptides in multiple samples without *a priori* knowledge of the identity of the molecular species.

Stable isotope labeling by amino acids in cell culture (SILAC) yields relative quantitative information (57). In the SILAC approach, one cell culture is grown in unlabeled medium while another is grown in medium with a stable heavy isotope, usually $C^{13}$ or $N^{15}$, such that one of the amino acids, such as lysine or arginine, is labeled. The cells are lysed and equal amounts of protein from the two samples are mixed together. The mixed samples are processed and analyzed using liquid chromatography coupled to tandem mass spectrometry. Cryptides can be identified by their fragmentation patterns, and relative quantitative information can be generated by comparing the integrated peak areas at the precursor levels for the labeled and unlabeled cryptides. The biggest advantage of SILAC is that experimental variables in processing of the samples are normalized, as the two samples for comparison are processed together after cell lysis. SILAC has been successfully used in a number of studies involving a wide range of processes (58–63).

SILAC was originally developed for bottom-up proteomics. More recently, it has been used also for top-down proteomics, and, as such, it is conceptually possible to determine the changes in the levels of cryptide precursors. Mathias Mann and co-workers initially demonstrated the applicability of SILAC-based quantitation in top-down proteomics (64), and later, David Muddiman and co-workers used SILAC to identify 11 intact proteins from human embryonic stem cells (65). These studies were designed only to test the possibility of identifying differentially expressed proteins with SILAC-based quantitation, but the results were promising for the future applications of this technology (64,65). It is possible to imagine determining changes in the levels of cryptide precursors using SILAC. Muddiman and co-workers also mathematically modeled the variability introduced by the biochemical processes of incorporation of labeled amino acids and metabolic conversion of arginines to prolines inside the cells during cell culture, information that will be very helpful for future top-down SILAC quantitation experiments (65).

Another isotope tagging-based quantitation method relies on tagging of the amino group by either deuterated (heavy) or protiated (light) acetic anhydride (66). Lloyd Fricker and co-workers have used this labeling approach to determine the changes in the brain peptidome under different condition and to identify differences in the levels of hemoglobin-derived peptides in blood, heart, and brain (18,66,67).

## CONCLUSION

There are a multitude of proteomic tools available that can be readily adapted to the study of cryptides. The strategic advantage of proteomic methods is that multiple cryptides can be assayed in a single experiment, which is not possible with any other approach. Discovery-mode cryptomic tools can be used to identify novel cryptides, and quantitative-mode tools can be used to build functional models for the action of the cryptide(s) of interest (Fig. 2). Adaptation of these tools is expected to lead to an explosion of information about the cryptome. An important challenge in utilizing mass spectrometry-based proteomics for cryptomic analysis is to establish the methodological paradigms for their routine use. We envision that the interaction between these two fields will not only benefit cryptomics but will also lead to the development of new analytical technologies that will benefit other areas of research.

## ACKNOWLEDGMENTS

We acknowledge Elizabeth M. Link for helpful comments and suggestions in the preparation of this manuscript. P.S. and A.J.L. were supported by NIH grant GM64779.

## REFERENCES

1. Tyers M, Mann M. From genomics to proteomics. Nature. 2003;422(6928):193–7. doi:10.1038/nature01510.
2. Autelitano DJ, Rajic A, Smith AI, Berndt MC, Ilag LL, Vadas M. The cryptome: a subset of the proteome, comprising cryptic peptides with distinct bioactivities. Drug Discov Today. 2006;11(7–8):306–14. doi:10.1016/j.drudis.2006.02.003.
3. Pimenta DC, Lebrun I. Cryptides: buried secrets in proteins. Peptides. 2007;28(12):2403–10. doi:10.1016/j.peptides.2007.10.005.
4. Ueki N, Someya K, Matsuo Y, Wakamatsu K, Mukai H. Cryptides: functional cryptic peptides hidden in protein structures. Pept Sci. 2007;88(2):190–8.

5. Fortini ME. Notch and presenilin: a proteolytic mechanism emerges. Curr Opin Cell Biol. 2001;13(5):627–34.
6. De Strooper B, Annaert W, Cupers P, Saftig P, Craessaerts K, Mumm JS, *et al.* A presenilin-1-dependent gamma-secretase-like protease mediates release of Notch intracellular domain. Nature. 1999;398(6727):518–22.
7. Mukai H, Hokari Y, Seki T, Takao T, Kubota M, Matsuo Y, *et al.* Discovery of mitocryptide-1, a neutrophil-activating cryptide from healthy porcine heart. J Biol Chem. 2008;283(45):30596–605.
8. Zania P, Gourni D, Aplin AC, Nicosia RF, Flordellis CS, Maragoudakis ME, *et al.* Parstatin, the cleaved peptide on proteinase-activated receptor 1 activation, is a potent inhibitor of angiogenesis. J Pharmacol Exp Ther. 2009;328(2):378–89.
9. Mains RE, Eipper BA, Ling N. Common precursor to corticotropins and endorphins. Proc Natl Acad Sci USA. 1977;74(7):3014–8.
10. Nakanishi S, Inoue A, Kita T, Nakamura M, Chang ACY, Cohen SN, *et al.* Nucleotide sequence of cloned cDNA for bovine corticotropin-[beta]-lipotropin precursor. Nature. 1979;278(5703):423–7. doi:10.1038/278423a0.
11. Roberts JL, Herbert E. Characterization of a common precursor to corticotropin and beta-lipotropin: cell-free synthesis of the precursor and identification of corticotropin peptides in the molecule. Proc Natl Acad Sci USA. 1977;74(11):4826–30.
12. Roberts JL, Herbert E. Characterization of a common precursor to corticotropin and beta-lipotropin: identification of beta-lipotropin peptides and their arrangement relative to corticotropin in the precursor synthesized in a cell-free system. Proc Natl Acad Sci USA. 1977;74(12):5300–4.
13. Herbert E. Discovery of pro-opiomelanocortin—a cellular polyprotein. Trends Biochem Sci. 1981;6:184–8.
14. Sousa JC, Berto RF, Gois EA, Fontenele-Cardi NC, Honorio JE Jr, Konno K, *et al.* Leptoglycin: a new glycine/leucine-rich antimicrobial peptide isolated from the skin secretion of the South American frog *Leptodactylus pentadactylus* (Leptodactylidae). Toxicon. 2009;54(1):23–32.
15. Nakatani S, Mano H, Sampei C, Shimizu J, Wada M. Chondroprotective effect of the bioactive peptide prolyl-hydroxyproline in mouse articular cartilage *in vitro* and *in vivo*. Osteoarthritis cartilage OARS Osteoarthritis Res Soc. 2009;17(12):1620–7.
16. Mukai H, Seki T, Nakano H, Hokari Y, Takao T, Kawanami M, *et al.* Mitocryptide-2: purification, identification, and characterization of a novel cryptide that activates neutrophils. J Immunol. 2009;182(8):5072–80.
17. Suder P, Nawrat D, Bielawski A, Zelek-Molik A, Raoof H, Dylag T, *et al.* Cryptic peptide derived from the rat neuropeptide FF precursor affects G-proteins linked to opioid receptors in the rat brain. Peptides. 2008;29(11):1988–93.
18. Gelman JS, Sironi J, Castro LM, Ferro ES, Fricker LD. Hemopressins and other hemoglobin-derived peptides in mouse brain: comparison between brain, blood, and heart peptidome and regulation in Cpefat/fat mice. J Neurochem. 2010;113(4):871–80.
19. Gomes I, Dale C, Casten K, Geigner M, Gozzo F, Ferro E, *et al.* Hemoglobin-derived peptides as novel type of bioactive signaling molecules. AAPS J. 2010;12:658–69.
20. O'Reilly MS, Holmgren L, Shing Y, Chen C, Rosenthal RA, Moses M, *et al.* Angiostatin: a novel angiogenesis inhibitor that mediates the suppression of metastases by a Lewis lung carcinoma. Cell. 1994;79(2):315–28.
21. O'Reilly MS, Boehm T, Shing Y, Fukai N, Vasios G, Lane WS, *et al.* Endostatin: an endogenous inhibitor of angiogenesis and tumor growth. Cell. 1997;88(2):277–85.
22. Deng LX, Pan XL, Wang Y, Wang LL, Zhou XE, Li M, *et al.* Hemoglobin and its derived peptides may play a role in the antibacterial mechanism of the vagina. Hum Reprod. 2009;24(1):211–8.
23. Liepke C, Baxmann S, Heine C, Breithaupt N, Standker L, Forssmann WG. Human hemoglobin-derived peptides exhibit antimicrobial activity: a class of host defense peptides. J Chromatogr B. 2003;791(1–2):345–56.
24. Dylag T, Pachuta A, Raoof H, Kotlinska J, Silberring J. A novel cryptic peptide derived from the rat neuropeptide FF precursor reverses antinociception and conditioned place preference induced by morphine. Peptides. 2008;29(3):473–8.
25. Jones S, Holm T, Mager I, Langel U, Howl J. Characterization of bioactive cell penetrating peptides from human cytochrome c: protein mimicry and the development of a novel apoptogenic agent. Chem Biol. 2010;17(7):735–44.
26. Ianzer D, Konno K, Xavier CH, Stocklin R, Santos RA, de Camargo AC, *et al.* Hemorphin and hemorphin-like peptides isolated from dog pancreas and sheep brain are able to potentiate bradykinin activity *in vivo*. Peptides. 2006;27(11):2957–66.
27. Nomizu M, Kim WH, Yamamura K, Utani A, Song S-Y, Otaka A, *et al.* Identification of cell binding sites in the laminin 1 chain carboxyl-terminal globular domain by systematic screening of synthetic peptides. J Biol Chem. 1995;270(35):20583–90.
28. Nomizu M, Kuratomi Y, Malinda KM, Song S-Y, Miyoshi K, Otaka A, *et al.* Cell binding sequences in mouse laminin α1 chain. J Biol Chem. 1998;273(49):32491–9.
29. Nomizu M, Kuratomi Y, Ponce ML, Song S-Y, Miyoshi K, Otaka A, *et al.* Cell adhesive sequences in mouse laminin [beta]1 chain. Arch Biochem Biophys. 2000;378(2):311–20.
30. Nomizu M, Kuratomi Y, Song S-Y, Ponce ML, Hoffman MP, Powell SK, *et al.* Identification of cell binding sequences in mouse laminin γ1 chain by systematic peptide screening. J Biol Chem. 1997;272(51):32198–205.
31. Franck J, Arafah K, Elayed M, Bonnel D, Vergara D, Jacquet A, *et al.* MALDI imaging mass spectrometry. Mol Cell Proteomics. 2009;8(9):2023–33.
32. Fleischer TC, Weaver CM, McAfee KJ, Jennings JL, Link AJ. Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes. Genes Dev. 2006;20(10):1294–307.
33. Link AJ. Multidimensional peptide separations in proteomics. Trends Biotechnol. 2002;20(12 Suppl):S8–13.
34. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, *et al.* Direct analysis of protein complexes using mass spectrometry. Nat Biotechnol. 1999;17(7):676–82.
35. Machtejevas E, Marko-Varga G, Lindberg C, Lubda D, Hendriks R, Unger KK. Profiling of endogenous peptides by multidimensional liquid chromatography: on-line automated sample cleanup for biomarker discovery in human urine. J Sep Sci. 2009;32(13):2223–32.
36. Clement CC, Cannizzo ES, Nastke MD, Sahu R, Olszewski W, Miller NE, *et al.* An expanded self-antigen peptidome is carried by the human lymph as compared to the plasma. PLoS One [Article]. 2010;5(3):e9863.
37. Petricoin EF, Belluco C, Araujo RP, Liotta LA. The blood peptidome: a higher dimension of information content for cancer biomarker discovery. Nat Rev Cancer. 2006;6(12):961–7. doi:10.1038/nrc2011.
38. Shen Y, Liu T, Tolić N, Petritis BO, Zhao R, Moore RJ, *et al.* Strategy for degradomic–peptidomic analysis of human blood plasma. J Proteome Res. 2010;9(5):2339–46.
39. Shen Y, Hixson KK, Tolić N, Camp DG, Purvine SO, Moore RJ, *et al.* Mass spectrometry analysis of proteome-wide proteolytic post-translational degradation of proteins. Anal Chem. 2008;80(15):5819–28.
40. Skold K, Svensson M, Kaplan A, Bjorkesten L, Astrom J, Andren PE. A neuroproteomic approach to targeting neuropeptides in the brain. Proteomics. 2002;2(4):447–54.
41. Hu L, Li X, Jiang X, Zhou H, Jiang X, Kong L, *et al.* Comprehensive peptidome analysis of mouse livers by size exclusion chromatography prefractionation and NanoLC-MS/MS identification. J Proteome Res. 2007;6(2):801–8.
42. Zheng X, Baker H, Hancock WS. Analysis of the low molecular weight serum peptidome using ultrafiltration and a hybrid ion trap-Fourier transform mass spectrometer. J Chromatogr A. 2006;1120(1–2):173–84.
43. Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. Anal Chem. 1997;69(23):4751–60.
44. Burnum KE, Frappier SL, Caprioli RM. Matrix-assisted laser desorption/ionization imaging mass spectrometry for the investigation of proteins and peptides. Annu Rev Anal Chem. 2008;1(1):689–705.

45. Millet LJ, Bora A, Sweedler JV, Gillette MU. Direct cellular peptidomics of supraoptic magnocellular and hippocampal neurons in low-density cocultures. ACS Chem Neurosci. 2009;1(1):36–48.
46. Rubakhin SS, Churchill JD, Greenough WT, Sweedler JV. Profiling signaling peptides in single mammalian cells using mass spectrometry. Anal Chem. 2006;78(20):7267–72.
47. Vaudel M, Sickmann A, Martens L. Peptide and protein quantification: a map of the minefield. Proteomics. 2010;10 (4):650–70.
48. Wilm M. Quantitative proteomics in biological research. Proteomics. 2009;9(20):4590–605.
49. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci USA. 2003;100 (12):6940–5.
50. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell. 2009;138(4):795–806.
51. Kirkpatrick DS, Gerber SA, Gygi SP. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. Methods. 2005;35 (3):265–73. doi:10.1016/j.ymeth.2004.08.018.
52. Kitteringham NR, Jenkins RE, Lane CS, Elliott VL, Park BK. Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics. J Chromatogr B. 2009;877 (13):1229–39. doi:10.1016/j.jchromb.2008.11.013.
53. Ross PL, Huang YLN, Marchese JN, Williamson B, Parker K, Hattan S, *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004;3(12):1154–69.
54. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003;75(8):1895–904.
55. Choe L, D'Ascenzo M, Relkin NR, Pappin D, Ross P, Williamson B, *et al.* 8-Plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. Proteomics. 2007;7(20):3651–60.
56. Dayon L, Hainard A, Licker V, Turck N, Kuhn K, Hochstrasser DF, *et al.* Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. Anal Chem. 2008;80(8):2921–31.
57. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1(5):376–86.
58. Mittler G, Butter F, Mann M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. Genome Res. 2009;19(2):284–93.
59. Pan C, Kumar C, Bohl S, Klingmueller U, Mann M. Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. Mol Cell Proteomics. 2009;8(3):443–50.
60. Ong S-E, Schenone M, Margolin AA, Li X, Do K, Doud MK, *et al.* Identifying the proteins to which small-molecule probes and drugs bind in cells. Proc Natl Acad Sci. 2009;106(12):4617–22.
61. Teckchandani A, Toida N, Goodchild J, Henderson C, Watts J, Wollscheid B, *et al.* Quantitative proteomics identifies a Dab2/integrin module regulating cell migration. J Cell Biol. 2009;186 (1):99–111.
62. Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, *et al.* Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science. 2009;325 (5942):834–40.
63. Jorgensen C, Sherman A, Chen GI, Pasculescu A, Poliakov A, Hsiung M, *et al.* Cell-specific information processing in segregating populations of Eph receptor ephrin-expressing cells. Science. 2009;326(5959):1502–9.
64. Waanders LF, Hanke S, Mann M. Top-down quantitation and characterization of SILAC-labeled proteins. J Am Soc Mass Spectrom. 2007;18(11):2058–64.
65. Collier TS, Sarkar P, Rao B, Muddiman DC. Quantitative top-down proteomics of SILAC labeled human embryonic stem cells. J Am Soc Mass Spectrom. 2010;21(6):879–89.
66. Che F-y, Fricker LD. Quantitation of neuropeptides in Cpefat/Cpefat mice using differential isotopic tags and mass spectrometry. Anal Chem. 2002;74(13):3190–8.
67. Che FY, Biswas R, Fricker LD. Relative quantitation of peptides in wild-type and Cpe(fat/fat) mouse pituitary using stable isotopic tags and mass spectrometry. J Mass Spectrom. 2005;40 (2):227–37.

# References


Aebersold, Ruedi, and Matthias Mann. 2003. “Mass Spectrometry-Based Proteomics.” *Nature* 422 (6928): 198–207. doi:10.1038/nature01511.

Åkerfelt, Malin, Richard I. Morimoto, and Lea Sistonen. 2010. “Heat Shock Factors: Integrators of Cell Stress, Development and Lifespan.” *Nature Reviews Molecular Cell Biology* 11 (8): 545–55. doi:10.1038/nrm2938.

Altelaar, A. F. Maarten, Javier Munoz, and Albert J. R. Heck. 2013. “Next-Generation Proteomics: Towards an Integrative View of Proteome Dynamics.” *Nature Reviews Genetics* 14 (1): 35–48. doi:10.1038/nrg3356.

Altenburger, Rolf, Thomas Backhaus, Wolfgang Boedeker, Michael Faust, and Martin Scholze. 2013. “Simplifying Complexity: Mixture Toxicity Assessment in the Last 20 Years.” *Environmental Toxicology and Chemistry* 32 (8): 1685–87. doi:10.1002/etc.2294.

Altenburger, Rolf, Monika Nendza, and Gerrit Schüürmann. 2003. “Mixture Toxicity and Its Modeling by Quantitative Structure-Activity Relationships.” *Environmental Toxicology and Chemistry* 22 (8): 1900–1915. doi:10.1897/01-386.

Altenburger, Rolf, Stefan Scholz, Mechthild Schmitt-Jansen, Wibke Busch, and Beate I. Escher. 2012. “Mixture Toxicity Revisited from a Toxicogenomic Perspective.” *Environmental Science & Technology* 46 (5): 2508–22. doi:10.1021/es2038036.

Altenburger, Rolf, Helge Walter, and Matthias Grote. 2004. “What Contributes to the Combined Effect of a Complex Mixture?” *Environmental Science & Technology* 38 (23): 6353–62. doi:10.1021/es049528k.

Amack, Jeffrey D., Aileen P. Paguio, and Mani S. Mahadevan. 1999. “Cis and Trans Effects of the Myotonic Dystrophy (DM) Mutation in a Cell Culture Model.” *Human Molecular Genetics* 8 (11): 1975–84. doi:10.1093/hmg/8.11.1975.

Amanatiadou, Elsa P., Giorgio L. Papadopoulos, John Strouboulis, and Ioannis S. Vizirianakis. 2015. “GATA1 and PU.1 Bind to Ribosomal Protein Genes in Erythroid Cells: Implications for Ribosomopathies.” *PLoS ONE* 10 (10): e0140077. doi:10.1371/journal.pone.0140077.

Amberg, David C., Dan Burke, and Jeffrey N. Strathern. 2005. *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*. CSHL Press.

Anant, Shrikant, Jeffrey O. Henderson, Debnath Mukhopadhyay, Naveenan Navaratnam, Susan Kennedy, Jing Min, and Nicholas O. Davidson. 2001. “Novel Role for RNA-Binding Protein CUGBP2 in Mammalian RNA Editing CUGBP2 MODULATES C TO U EDITING OF APOLIPOPROTEIN B mRNA BY INTERACTING WITH APOBEC-1 AND ACF, THE APOBEC-1 COMPLEMENTATION FACTOR.” *Journal of Biological Chemistry* 276 (50): 47338–51. doi:10.1074/jbc.M104911200.

Andersen, Gregers R, Poul Nissen, and Jens Nyborg. 2003. “Elongation Factors in Protein Biosynthesis.” *Trends in Biochemical Sciences* 28 (8): 434–41. doi:10.1016/S0968-0004(03)00162-2.

Andersen, Peter M., Katherine B. Sims, Winnie W. Xin, Rosemary Kiely, Gilmore O’Neill, John Ravits, Erik Pioro, et al. 2003. “Sixteen Novel Mutations in the Cu/Zn Superoxide Dismutase Gene in Amyotrophic Lateral Sclerosis: A Decade of Discoveries, Defects and Disputes.” *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders: Official Publication of the World Federation of Neurology, Research Group on Motor Neuron Diseases* 4 (2): 62–73.

Anderson, Paul, and Nancy Kedersha. 2009. “RNA Granules: Post-Transcriptional and Epigenetic Modulators of Gene Expression.” *Nature Reviews Molecular Cell Biology* 10 (6): 430–36. doi:10.1038/nrm2694.

Anderson, W. Marshall, Anette Grundholm, and Bruce H. Sells. 1975. “Modification of Ribosomal Proteins during Liver Regeneration.” *Biochemical and Biophysical Research Communications* 62 (3): 669–76. doi:10.1016/0006-291X(75)90451-9.

Ankomah, Peter, and Bruce R. Levin. 2012. “Two-Drug Antimicrobial Chemotherapy: A Mathematical Model and Experiments with Mycobacterium Marinum.” *PLoS Pathog* 8 (1): e1002487. doi:10.1371/journal.ppat.1002487.

Apel, Klaus, and Heribert Hirt. 2004. “REACTIVE OXYGEN SPECIES: Metabolism, Oxidative Stress, and Signal Transduction.” *Annual Review of Plant Biology* 55 (1): 373–99. doi:10.1146/annurev.arplant.55.031903.141701.

Arnold, Randy J., Bogdan Polevoda, James P. Reilly, and Fred Sherman. 1999. “The Action of N-Terminal Acetyltransferases on Yeast Ribosomal Proteins.” *Journal of Biological Chemistry* 274 (52): 37035–40. doi:10.1074/jbc.274.52.37035.

Arragain, Simon, Ricardo Garcia-Serres, Geneviève Blondin, Thierry Douki, Martin Clemancey, Jean-Marc Latour, Farhad Forouhar, et al. 2010. “Post-Translational Modification of Ribosomal Proteins STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF RimO FROM THERMOTOGA MARITIMA, A RADICAL S-ADENOSYLMETHIONINE METHYLTHIOTRANSFERASE.” *Journal of Biological Chemistry* 285 (8): 5792–5801. doi:10.1074/jbc.M109.065516.

Artero, Ruben, Andreas Prokop, Nuria Paricio, Gerrit Begemann, Ignacio Pueyo, Marek Mlodzik, Manuel Perez-Alonso, and Mary K. Baylies. 1998. “The Muscleblind Gene Participates in the Organization of Z-Bands and Epidermal Attachments ofDrosophilaMuscles and Is Regulated byDmef2.” *Developmental Biology* 195 (2): 131–43. doi:10.1006/dbio.1997.8833.

Ashizawa, Tetsuo, and HenryF. Epstein. 1991. “Ethnic Distribution of Myotonic Dystrophy Gene.” *The Lancet* 338 (8767): 642–43. doi:10.1016/0140-6736(91)90659-D.

Aslanidis, Charalampos, Gert Jansen, Chris Amemiya, Gary Shutler, Mani Mahadevan, Catherine Tsilfidis, Chira Chen, et al. 1992. “Cloning of the Essential Myotonic Dystrophy Region and Mapping of the Putative Defect.” *Nature* 355 (6360): 548–51. doi:10.1038/355548a0.

Aspesi, Anna, Elisa Pavesi, Elisa Robotti, Rossella Crescitelli, Ilenia Boria, Federica Avondo, Hélène Moniz, et al. 2014. “Dissecting the Transcriptional Phenotype of Ribosomal Protein Deficiency: Implications for Diamond-Blackfan Anemia.” *Gene* 545 (2): 282–89. doi:10.1016/j.gene.2014.04.077.

Bachinski, L. L., T. Czernuszewicz, L. S. Ramagli, T. Suominen, M. D. Shriver, B. Udd, M. J. Siciliano, and R. Krahe. 2009. “Premutation Allele Pool in Myotonic Dystrophy Type 2.” *Neurology* 72 (6): 490–97. doi:10.1212/01.wnl.0000333665.01888.33.

Baker Brachmann, Carrie, Adrian Davies, Gregory J. Cost, Emerita Caputo, Joachim Li, Philip Hieter, and Jef D. Boeke. 1998. “Designer Deletion Strains Derived from Saccharomyces Cerevisiae S288C: A Useful Set of Strains and Plasmids for PCR-Mediated Gene Disruption and Other Applications.” *Yeast* 14 (2): 115–32. doi:10.1002/(SICI)1097-0061(19980130)14:2<115::AID-YEA204>3.0.CO;2-2.

Bantscheff, Marcus, Simone Lemeer, Mikhail M. Savitski, and Bernhard Kuster. 2012. “Quantitative Mass Spectrometry in Proteomics: Critical Review Update from 2007 to the Present.” *Analytical and Bioanalytical Chemistry* 404 (4): 939–65. doi:10.1007/s00216-012-6203-4.

Bantscheff, Marcus, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. 2007. “Quantitative Mass Spectrometry in Proteomics: A Critical Review.” *Analytical and Bioanalytical Chemistry* 389 (4): 1017–31. doi:10.1007/s00216-007-1486-6.

Bateson, William. 1909. *Mendel's Principles of Heredity, by W. Bateson*. Cambridge [Eng.] University Press. http://archive.org/details/mendelsprinciple00bate.

Begemann, G., N. Paricio, R. Artero, I. Kiss, M. Perez-Alonso, and M. Mlodzik. 1997. "Muscleblind, a Gene Required for Photoreceptor Differentiation in Drosophila, Encodes Novel Nuclear Cys3His-Type Zinc-Finger-Containing Proteins." *Development* 124 (21): 4321–31.

Belden, Jason B, Robert J Gilliom, and Michael J Lydy. 2007. "How Well Can We Predict the Toxicity of Pesticide Mixtures to Aquatic Life?" *Integrated Environmental Assessment and Management* 3 (3): e1–5. doi:10.1002/ieam.5630030326.

Beltrao, Pedro, Peer Bork, Nevan J. Krogan, and Vera van Noort. 2013. "Evolution and Functional Cross-Talk of Protein Post-Translational Modifications." *Molecular Systems Biology* 9 (1): n/a – n/a. doi:10.1002/msb.201304521.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Berenbaum, M. C. 1989. "What Is Synergy?" *Pharmacological Reviews* 41 (2): 93–141.

Berg, Robert A. van den, Huub CJ Hoefsloot, Johan A. Westerhuis, Age K. Smilde, and Mariët J. van der Werf. 2006. "Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data." *BMC Genomics* 7 (1): 142. doi:10.1186/1471-2164-7-142.

Bhagwati, Satyakam, S. A. Shafiq, and Weimin Xu. 1999. "(CTG)n Repeats Markedly Inhibit Differentiation of the C2C12 Myoblast Cell Line: Implications for Congenital Myotonic Dystrophy." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1453 (2): 221–29. doi:10.1016/S0925-4439(98)00104-5.

Bianchi, Marzia, Elisa Giacomini, Rita Crinelli, Lucia Radici, Elisa Carloni, and Mauro Magnani. 2015. "Dynamic Transcription of Ubiquitin Genes under Basal and Stressful Conditions and New Insights into the Multiple UBC Transcript Variants." *Gene* 573 (1): 100–109. doi:10.1016/j.gene.2015.07.030.

Blokhina, Olga, Eija Virolainen, and Kurt V. Fagerstedt. 2003. "Antioxidants, Oxidative Damage and Oxygen Deprivation Stress: A Review." *Annals of Botany* 91 (2): 179–94. doi:10.1093/aob/mcf118.

Boehringer, Daniel, Rolf Thermann, Antje Ostareck-Lederer, Joe D. Lewis, and Holger Stark. 2005. "Structure of the Hepatitis C Virus IRES Bound to the Human 80S Ribosome: Remodeling of the HCV IRES." *Structure* 13 (11): 1695–1706. doi:10.1016/j.str.2005.08.008.

Boisvert, François-Michel, Silvana van Koningsbruggen, Joaquín Navascués, and Angus I. Lamond. 2007. "The Multifunctional Nucleolus." *Nature Reviews Molecular Cell Biology* 8 (7): 574–85. doi:10.1038/nrm2184.

Brauer, Matthew J., Curtis Huttenhower, Edoardo M. Airoldi, Rachel Rosenstein, John C. Matese, David Gresham, Viktor M. Boer, Olga G. Troyanskaya, and David Botstein. 2008. "Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast." *Molecular Biology of the Cell* 19 (1): 352–67. doi:10.1091/mbc.E07-08-0779.

Breker, Michal, and Maya Schuldiner. 2014. "The Emergence of Proteome-Wide Technologies: Systematic Analysis of Proteins Comes of Age." *Nature Reviews Molecular Cell Biology* 15 (7): 453–64. doi:10.1038/nrm3821.

Breslow, David K., Dale M. Cameron, Sean R. Collins, Maya Schuldiner, Jacob Stewart-Ornstein, Heather W. Newman, Sigurd Braun, Hiten D. Madhani, Nevan J. Krogan, and Jonathan S. Weissman. 2008. "A Comprehensive Strategy Enabling High-Resolution Functional Analysis of the Yeast Genome." *Nature Methods* 5 (8): 711–18. doi:10.1038/nmeth.1234.

Brisson, Diane, Marie-Claude Vohl, Julie St-Pierre, Thomas J. Hudson, and Daniel Gaudet. 2001. “Glycerol: A Neglected Variable in Metabolic Processes?” *BioEssays* 23 (6): 534–42. doi:10.1002/bies.1073.

Brook, J. David, Mila E. McCurrach, Helen G. Harley, Alan J. Buckler, Deanna Church, Hiroyuki Aburatani, Kent Hunter, et al. 1992. “Molecular Basis of Myotonic Dystrophy: Expansion of a Trinucleotide (CTG) Repeat at the 3′ End of a Transcript Encoding a Protein Kinase Family Member.” *Cell* 68 (4): 799–808. doi:10.1016/0092-8674(92)90154-5.

Brooks, Susan S., Alissa L. Wall, Christelle Golzio, David W. Reid, Amalia Kondyles, Jason R. Willer, Christina Botti, Christopher V. Nicchitta, Nicholas Katsanis, and Erica E. Davis. 2014. “A Novel Ribosomopathy Caused by Dysfunction of RPL10 Disrupts Neurodevelopment and Causes X-Linked Microcephaly in Humans.” *Genetics* 198 (2): 723–33. doi:10.1534/genetics.114.168211.

Browne, Christopher M., Parimal Samir, J. Scott Fites, Seth A. Villarreal, and Andrew J. Link. 2013. “The Yeast Eukaryotic Translation Initiation Factor 2B Translation Initiation Complex Interacts with the Fatty Acid Synthesis Enzyme YBR159W and Endoplasmic Reticulum Membranes.” *Molecular and Cellular Biology* 33 (5): 1041–56. doi:10.1128/MCB.00811-12.

Brunner, H. G., H. Smeets, H. M. M. Lambermon, M. Coerwinkel-Driessen, B. A. van Oost, B. Wieringa, and H. H. Ropers. 1989. “A Multipoint Linkage Map around the Locus for Myotonic Dystrophy on Chromosome 19.” *Genomics* 5 (3): 589–95. doi:10.1016/0888-7543(89)90027-X.

Cagney, Gerard, and Andrew Emili. 2002. “De Novo Peptide Sequencing and Quantitative Profiling of Complex Protein Mixtures Using Mass-Coded Abundance Tagging.” *Nature Biotechnology* 20 (2): 163–70. doi:10.1038/nbt0202-163.

Cardani, Rosanna, Enrico Bugiardini, Laura V. Renna, Giulia Rossi, Graziano Colombo, Rea Valaperta, Giuseppe Novelli, Annalisa Botta, and Giovanni Meola. 2013. “Overexpression of CUGBP1 in Skeletal Muscle from Adult Classic Myotonic Dystrophy Type 1 but Not from Myotonic Dystrophy Type 2.” *PLoS ONE* 8 (12): e83777. doi:10.1371/journal.pone.0083777.

Carvalho, Claudine M., Anésia A. Santos, Silvana R. Pires, Carolina S. Rocha, Daniela I. Saraiva, João Paulo B. Machado, Eliciane C. Mattos, Luciano G. Fietto, and Elizabeth P. B. Fontes. 2008. “Regulated Nuclear Trafficking of rpL10A Mediated by NIK1 Represents a Defense Strategy of Plant Cells against Virus.” *PLoS Pathog* 4 (12): e1000247. doi:10.1371/journal.ppat.1000247.

Chan, Clement T. Y., Yan Ling Joy Pang, Wenjun Deng, I. Ramesh Babu, Madhu Dyavaiah, Thomas J. Begley, and Peter C. Dedon. 2012. “Reprogramming of tRNA Modifications Controls the Oxidative Stress Response by Codon-Biased Translation of Proteins.” *Nature Communications* 3 (July): 937. doi:10.1038/ncomms1938.

Chapin III, F. Stuart, Kellar Autumn, and Francisco Pugnaire. 1993. “Evolution of Suites of Traits in Response to Environmental Stress.” *The American Naturalist* 142 (July): S78–92.

Chappell, Stephen A., Gerald M. Edelman, and Vincent P. Mauro. 2000. “A 9-Nt Segment of a Cellular mRNA Can Function as an Internal Ribosome Entry Site (IRES) and When Present in Linked Multiple Copies Greatly Enhances IRES Activity.” *Proceedings of the National Academy of Sciences* 97 (4): 1536–41. doi:10.1073/pnas.97.4.1536.

Charlet-B., Nicolas, Rajesh S. Savkur, Gopal Singh, Anne V. Philips, Elizabeth A. Grice, and Thomas A. Cooper. 2002. “Loss of the Muscle-Specific Chloride Channel in Type 1 Myotonic Dystrophy Due to Misregulated Alternative Splicing.” *Molecular Cell* 10 (1): 45–53. doi:10.1016/S1097-2765(02)00572-5.

Chen, Baozhi, Michael E. Dodge, Wei Tang, Jianming Lu, Zhiqiang Ma, Chih-Wei Fan, Shuguang Wei, et al. 2009. “Small Molecule–mediated Disruption of Wnt-Dependent

Signaling in Tissue Regeneration and Cancer." *Nature Chemical Biology* 5 (2): 100–107. doi:10.1038/nchembio.137.
Chen, Bo, Bingxin Shen, and Joachim Frank. 2014. "Particle Migration Analysis in Iterative Classification of Cryo-EM Single-Particle Data." *Journal of Structural Biology* 188 (3): 267–73. doi:10.1016/j.jsb.2014.10.006.
Chen, Wei, Yucheng Wang, Yoko Abe, Lukas Cheney, Bjarne Udd, and Yi-Ping Li. 2007. "Haploinsuffciency for Znf9 in Znf9+/− Mice Is Associated with Multiorgan Abnormalities Resembling Myotonic Dystrophy." *Journal of Molecular Biology* 368 (1): 8–17. doi:10.1016/j.jmb.2007.01.088.
Chen, Wenqiong J., and Tong Zhu. 2004. "Networks of Transcription Factors with Roles in Environmental Stress Response." *Trends in Plant Science* 9 (12): 591–96. doi:10.1016/j.tplants.2004.10.007.
Cherry, J. Michael, Caroline Adler, Catherine Ball, Stephen A. Chervitz, Selina S. Dwight, Erich T. Hester, Yankai Jia, et al. 1998. "SGD: Saccharomyces Genome Database." *Nucleic Acids Research* 26 (1): 73–79. doi:10.1093/nar/26.1.73.
Cho, Diane H., and Stephen J. Tapscott. 2007. "Myotonic Dystrophy: Emerging Mechanisms for DM1 and DM2." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, The Muscular Dystrophies: Molecular Basis and Therapeutic Strategies, 1772 (2): 195–204. doi:10.1016/j.bbadis.2006.05.013.
Choudhary, Chunaram, and Matthias Mann. 2010. "Decoding Signalling Networks by Mass Spectrometry-Based Proteomics." *Nature Reviews Molecular Cell Biology* 11 (6): 427–39. doi:10.1038/nrm2900.
Christie, Karen R., Shuai Weng, Rama Balakrishnan, Maria C. Costanzo, Kara Dolinski, Selina S. Dwight, Stacia R. Engel, et al. 2004. "Saccharomyces Genome Database (SGD) Provides Tools to Identify and Analyze Sequences from Saccharomyces Cerevisiae and Related Sequences from Other Organisms." *Nucleic Acids Research* 32 (suppl 1): D311–14. doi:10.1093/nar/gkh033.
CHUNG, LELAND W. K., ADAM BASEMAN, VASILY ASSIKIS, and HAIYEN E. ZHAU. 2005. "MOLECULAR INSIGHTS INTO PROSTATE CANCER PROGRESSION: THE MISSING LINK OF TUMOR MICROENVIRONMENT." *The Journal of Urology* 173 (1): 10–20. doi:10.1097/01.ju.0000141582.15218.10.
Clancy, Trevor, and Eivind Hovig. 2014. "From Proteomes to Complexomes in the Era of Systems Biology." *PROTEOMICS* 14 (1): 24–41. doi:10.1002/pmic.201300230.
Clarke, Sally RmcK, Megan Barnden, Christian Kurts, Francis R. Carbone, Jacques Fap Miller, and William R. Heath. 2000. "Characterization of the Ovalbumin-Specific TCR Transgenic Line OT-I: MHC Elements for Positive and Negative Selection." *Immunology and Cell Biology* 78 (2): 110–17. doi:10.1046/j.1440-1711.2000.00889.x.
Clauset, A., C. Shalizi, and M. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51 (4): 661–703. doi:10.1137/070710111.
Cloonan, Nicole, Alistair R. R. Forrest, Gabriel Kolle, Brooke B. A. Gardiner, Geoffrey J. Faulkner, Mellissa K. Brown, Darrin F. Taylor, et al. 2008. "Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing." *Nature Methods* 5 (7): 613–19. doi:10.1038/nmeth.1223.
Cordell, Heather J. 2002. "Epistasis: What It Means, What It Doesn't Mean, and Statistical Methods to Detect It in Humans." *Human Molecular Genetics* 11 (20): 2463–68. doi:10.1093/hmg/11.20.2463.
———. 2009. "Detecting Gene–gene Interactions That Underlie Human Diseases." *Nature Reviews Genetics* 10 (6): 392–404. doi:10.1038/nrg2579.
Crick, Francis. 1970. "Central Dogma of Molecular Biology." *Nature* 227 (5258): 561–63. doi:10.1038/227561a0.

Cunningham, Fiona, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, et al. 2015. “Ensembl 2015.” *Nucleic Acids Research* 43 (D1): D662–69. doi:10.1093/nar/gku1010.

Danilova, Nadia, and Hanna T. Gazda. 2015. “Ribosomopathies: How a Common Root Can Cause a Tree of Pathologies.” *Disease Models & Mechanisms* 8 (9): 1013–26. doi:10.1242/dmm.020529.

Day, J. W., K. Ricker, J. F. Jacobsen, L. J. Rasmussen, K. A. Dick, W. Kress, C. Schneider, et al. 2003. “Myotonic Dystrophy Type 2 Molecular, Diagnostic and Clinical Spectrum.” *Neurology* 60 (4): 657–64. doi:10.1212/01.WNL.0000054481.84978.F9.

De Angelis, Maria, and Marco Gobbetti. 2004. “Environmental Stress Responses in Lactobacillus: A Review.” *PROTEOMICS* 4 (1): 106–22. doi:10.1002/pmic.200300497.

de la Cruz, Jesus, Katrin Karbstein, and John L. Woolford Jr. 2015. “Functions of Ribosomal Proteins in Assembly of Eukaryotic Ribosomes In Vivo.” *Annual Review of Biochemistry* 84 (1): null. doi:10.1146/annurev-biochem-060614-033917.

de León, Mario Bermúdez, and Bulmaro Cisneros. 2008. “Myotonic Dystrophy 1 in the Nervous System: From the Clinic to Molecular Mechanisms.” *Journal of Neuroscience Research* 86 (1): 18–26. doi:10.1002/jnr.21377.

de Ligt, Joep, Marjolein H. Willemsen, Bregje W.M. van Bon, Tjitske Kleefstra, Helger G. Yntema, Thessa Kroes, Anneke T. Vulto-van Silfhout, et al. 2012. “Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability.” *New England Journal of Medicine* 367 (20): 1921–29. doi:10.1056/NEJMoa1206524.

de Munain, A. López, A. M. Cobo, E. Huguet, J. F. Martí Massó, K. Johnson, and M. Baiget. 1994. “CTG Trinucleotide Repeat Variability in Identical Twins with Myotonic Dystrophy.” *Annals of Neurology* 35 (3): 374–75. doi:10.1002/ana.410350323.

de Nadal, Eulàlia, Gustav Ammerer, and Francesc Posas. 2011. “Controlling Gene Expression in Response to Stress.” *Nature Reviews Genetics* 12 (12): 833–45. doi:10.1038/nrg3055.

Deneer, John W. 2000. “Toxicity of Mixtures of Pesticides in Aquatic Systems.” *Pest Management Science* 56 (6): 516–20. doi:10.1002/(SICI)1526-4998(200006)56:6<516::AID-PS163>3.0.CO;2-0.

Dere, Ruhee, and Robert D. Wells. 2006. “DM2 CCTG•CAGG Repeats Are Crossover Hotspots That Are More Prone to Expansions than the DM1 CTG•CAG Repeats in Escherichia Coli.” *Journal of Molecular Biology* 360 (1): 21–36. doi:10.1016/j.jmb.2006.05.012.

DeRisi, Joseph L., Vishwanath R. Iyer, and Patrick O. Brown. 1997. “Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale.” *Science* 278 (5338): 680–86. doi:10.1126/science.278.5338.680.

DeRisi, J., L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. 1996. “Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer.” *Nature Genetics* 14 (4): 457–60. doi:10.1038/ng1296-457.

Dever, Thomas E., and Rachel Green. 2012. “The Elongation, Termination, and Recycling Phases of Translation in Eukaryotes.” *Cold Spring Harbor Perspectives in Biology* 4 (7): a013706. doi:10.1101/cshperspect.a013706.

Dieterle, Frank, Alfred Ross, Götz Schlotterbeck, and Hans Senn. 2006. “Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics.” *Analytical Chemistry* 78 (13): 4281–90. doi:10.1021/ac051632c.

Dixon, Scott J., Michael Costanzo, Anastasia Baryshnikova, Brenda Andrews, and Charles Boone. 2009. “Systematic Mapping of Genetic Interaction Networks.” *Annual Review of Genetics* 43 (1): 601–25. doi:10.1146/annurev.genet.39.073003.114751.

Douglas, Andrew G. L., and Matthew J. A. Wood. 2011. “RNA Splicing: Disease and Therapy.” *Briefings in Functional Genomics* 10 (3): 151–64. doi:10.1093/bfgp/elr020.

Draptchinskaia, Natalia, Peter Gustavsson, Björn Andersson, Monica Pettersson, Thiébaut-Willig, Irma Dianzani, Sarah Ball, et al. 1999. “The Gene Encoding Ribosomal Protein S19 Is Mutated in Diamond-Blackfan Anaemia.” *Nature Genetics* 21 (2): 169–75. doi:10.1038/5951.

Drissi, Romain, Marie-Line Dubois, and François-Michel Boisvert. 2013. “Proteomics Methods for Subcellular Proteome Analysis.” *FEBS Journal* 280 (22): 5626–34. doi:10.1111/febs.12502.

Du, Hongqing, Melissa S. Cline, Robert J. Osborne, Daniel L. Tuttle, Tyson A. Clark, John Paul Donohue, Megan P. Hall, et al. 2010. “Aberrant Alternative Splicing and Extracellular Matrix Gene Expression in Mouse Models of Myotonic Dystrophy.” *Nature Structural & Molecular Biology* 17 (2): 187–93. doi:10.1038/nsmb.1720.

Duncan, Roger, and Edwin H. McCONKEY. 1982. “Preferential Utilization of Phosphorylated 40-S Ribosomal Subunits during Initiation Complex Formation.” *European Journal of Biochemistry* 123 (3): 535–38. doi:10.1111/j.1432-1033.1982.tb06564.x.

Dunn, Olive Jean. 1961. “Multiple Comparisons among Means.” *Journal of the American Statistical Association* 56 (293): 52–64. doi:10.1080/01621459.1961.10482090.

Durbin, B. P., J. S. Hardin, D. M. Hawkins, and D. M. Rocke. 2002. “A Variance-Stabilizing Transformation for Gene-Expression Microarray Data.” *Bioinformatics* 18 (suppl 1): S105–10. doi:10.1093/bioinformatics/18.suppl_1.S105.

Durrani, O. M., N. N. Tehrani, J. E. Marr, P. Moradi, P. Stavrou, and P. I. Murray. 2004. “Degree, Duration, and Causes of Visual Loss in Uveitis.” *British Journal of Ophthalmology* 88 (9): 1159–62. doi:10.1136/bjo.2003.037226.

Dutta, Bhaskar, Harin Kanani, John Quackenbush, and Maria I. Klapa. 2009. “Time-Series Integrated ‘omic’ Analyses to Elucidate Short-Term Stress-Induced Responses in Plant Liquid Cultures.” *Biotechnology and Bioengineering* 102 (1): 264–79. doi:10.1002/bit.22036.

Eisen, Michael B., Paul T. Spellman, Patrick O. Brown, and David Botstein. 1998. “Cluster Analysis and Display of Genome-Wide Expression Patterns.” *Proceedings of the National Academy of Sciences* 95 (25): 14863–68.

Eng, Jimmy K., Bernd Fischer, Jonas Grossmann, and Michael J. MacCoss. 2008. “A Fast SEQUEST Cross Correlation Algorithm.” *Journal of Proteome Research* 7 (10): 4598–4602. doi:10.1021/pr800420s.

Eng, Jimmy K., Ashley L. McCormack, and John R. Yates. 1994. “An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.” *Journal of the American Society for Mass Spectrometry* 5 (11): 976–89. doi:10.1016/1044-0305(94)80016-2.

Everley, Patrick A., Jeroen Krijgsveld, Bruce R. Zetter, and Steven P. Gygi. 2004. “Quantitative Cancer Proteomics: Stable Isotope Labeling with Amino Acids in Cell Culture (SILAC) as a Tool for Prostate Cancer Research.” *Molecular & Cellular Proteomics* 3 (7): 729–35. doi:10.1074/mcp.M400021-MCP200.

Fan, Mingsheng, Jianbo Shen, Lixing Yuan, Rongfeng Jiang, Xinping Chen, William J. Davies, and Fusuo Zhang. 2011. “Improving Crop Productivity and Resource Use Efficiency to Ensure Food Security and Environmental Quality in China.” *Journal of Experimental Botany*, September, err248. doi:10.1093/jxb/err248.

Fardaei, Majid, Ken Larkin, J. David Brook, and Marion G. Hamshere. 2001. “In Vivo Co-Localisation of MBNL Protein with DMPK Expanded-Repeat Transcripts.” *Nucleic Acids Research* 29 (13): 2766–71. doi:10.1093/nar/29.13.2766.

Fardaei, Majid, Mark T. Rogers, Helena M. Thorpe, Kenneth Larkin, Marion G. Hamshere, Peter S. Harper, and J. David Brook. 2002. “Three Proteins, MBNL, MBLL and MBXL, Co-Localize in Vivo with Nuclear Foci of Expanded-Repeat Transcripts in DM1 and DM2 Cells.” *Human Molecular Genetics* 11 (7): 805–14. doi:10.1093/hmg/11.7.805.

Faust, M., R. Altenburger, T. Backhaus, H. Blanck, W. Boedeker, P. Gramatica, V. Hamer, M. Scholze, M. Vighi, and L. H. Grimme. 2001. “Predicting the Joint Algal Toxicity of Multi-Component S-Triazine Mixtures at Low-Effect Concentrations of Individual Toxicants.” *Aquatic Toxicology* 56 (1): 13–32. doi:10.1016/S0166-445X(01)00187-4.

Feder, Martin E., and Gretchen E. Hofmann. 1999. “Heat-Shock Proteins, Molecular Chaperones, and the Stress Response: Evolutionary and Ecological Physiology.” *Annual Review of Physiology* 61 (1): 243–82. doi:10.1146/annurev.physiol.61.1.243.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. 1989. “Electrospray Ionization for Mass Spectrometry of Large Biomolecules.” *Science* 246 (4926): 64–71. doi:10.1126/science.2675315.

Fernández, Israel S., Xiao-Chen Bai, Garib Murshudov, Sjors H. W. Scheres, and V. Ramakrishnan. 2014. “Initiation of Translation by Cricket Paralysis Virus IRES Requires Its Translocation in the Ribosome.” *Cell* 157 (4): 823–31. doi:10.1016/j.cell.2014.04.015.

Finger, Elizabeth C., and Amato J. Giaccia. 2010. “Hypoxia, Inflammation, and the Tumor Microenvironment in Metastatic Disease.” *Cancer and Metastasis Reviews* 29 (2): 285–93. doi:10.1007/s10555-010-9224-5.

Firth, Andrew E., and Ian Brierley. 2012. “Non-Canonical Translation in RNA Viruses.” *Journal of General Virology* 93 (Pt 7): 1385–1409. doi:10.1099/vir.0.042499-0.

Fisher, R. A. 1958. *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. OUP Oxford.

Franckenberg, Sibylle, Thomas Becker, and Roland Beckmann. 2012. “Structural View on Recycling of Archaeal and Eukaryotic Ribosomes after Canonical Termination and Ribosome Rescue.” *Current Opinion in Structural Biology*, Catalysis and regulation • Proteins, 22 (6): 786–96. doi:10.1016/j.sbi.2012.08.002.

Frank, Joachim. 2012. “Intermediate States during mRNA–tRNA Translocation.” *Current Opinion in Structural Biology*, Catalysis and regulation • Proteins, 22 (6): 778–85. doi:10.1016/j.sbi.2012.08.001.

Fu, Y. H., A. Pizzuti, R. G. Fenwick, J. King, S. Rajnarayan, P. W. Dunne, J. Dubel, G. A. Nasser, T. Ashizawa, and P. de Jong. 1992. “An Unstable Triplet Repeat in a Gene Related to Myotonic Muscular Dystrophy.” *Science (New York, N.Y.)* 255 (5049): 1256–58.

Fu, Ying-Hui, David L. Friedman, Stephen Richards, Joel A. Pearlman, Richard A. Gibbs, Antonio Pizzuti, Tetsuo Ashizawa, et al. 1993. “Decreased Expression of Myotonin-Protein Kinase Messenger RNA and Protein in Adult Form of Myotonic Dystrophy.” *Science*, New Series, 260 (5105): 235–38.

Gajadhar, Aaron S, and Forest M White. 2014. “System Level Dynamics of Post-Translational Modifications.” *Current Opinion in Biotechnology*, Nanobiotechnology • Systems biology, 28 (August): 83–87. doi:10.1016/j.copbio.2013.12.009.

Gamalinda, Michael, Uli Ohmayer, Jelena Jakovljevic, Beril Kumcuoglu, Joshua Woolford, Bertrade Mbom, Lawrence Lin, and John L. Woolford. 2014. “A Hierarchical Model for Assembly of Eukaryotic 60S Ribosomal Subunit Domains.” *Genes & Development* 28 (2): 198–210. doi:10.1101/gad.228825.113.

Gao, Hong, Julie M. Granka, and Marcus W. Feldman. 2010. “On the Classification of Epistatic Interactions.” *Genetics* 184 (3): 827–37. doi:10.1534/genetics.109.111120.

GARDNER, SHEA N. 2002. “Modeling Multi-Drug Chemotherapy: Tailoring Treatment to Individuals.” *Journal of Theoretical Biology* 214 (2): 181–207. doi:10.1006/jtbi.2001.2459.

Gasch, Audrey P., Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. 2000. “Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes.” *Molecular Biology of the Cell* 11 (12): 4241–57. doi:10.1091/mbc.11.12.4241.

Geiger, Tamar, Juergen Cox, Pawel Ostasiewicz, Jacek R. Wisniewski, and Matthias Mann. 2010. “Super-SILAC Mix for Quantitative Proteomics of Human Tumor Tissue.” *Nature Methods* 7 (5): 383–85. doi:10.1038/nmeth.1446.

Georges, Amédée des, Yaser Hashem, Anett Unbehaun, Robert A. Grassucci, Derek Taylor, Christopher U. T. Hellen, Tatyana V. Pestova, and Joachim Frank. 2014. “Structure of the Mammalian Ribosomal Pre-Termination Complex Associated with eRF1•eRF3•GDPNP.” *Nucleic Acids Research* 42 (5): 3409–18. doi:10.1093/nar/gkt1279.

Gerard, Jean-François, Michel Vancassel, and Brigitte Laffort. 1993. “Spread of Phenotypic Plasticity or Genetic Assimilation: The Possible Role of Genetic Constraints.” *Journal of Theoretical Biology* 164 (3): 341–49. doi:10.1006/jtbi.1993.1158.

Gerbasi, Vincent R., and Andrew J. Link. 2007. “The Myotonic Dystrophy Type 2 Protein ZNF9 Is Part of an ITAF Complex That Promotes Cap-Independent Translation.” *Molecular & Cellular Proteomics* 6 (6): 1049–58. doi:10.1074/mcp.M600384-MCP200.

Gerner, Christopher, Susanne Vejda, Dieter Gelbmann, Editha Bayer, Josef Gotzmann, Rolf Schulte-Hermann, and Wolfgang Mikulits. 2002. “Concomitant Determination of Absolute Values of Cellular Protein Amounts, Synthesis Rates, and Turnover Rates by Quantitative Proteome Profiling.” *Molecular & Cellular Proteomics* 1 (7): 528–37. doi:10.1074/mcp.M200026-MCP200.

Giaever, Guri, Angela M. Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, et al. 2002. “Functional Profiling of the Saccharomyces Cerevisiae Genome.” *Nature* 418 (6896): 387–91. doi:10.1038/nature00935.

Gingras, Anne-Claude, Brian Raught, and Nahum Sonenberg. 1999. “eIF4 Initiation Factors: Effectors of mRNA Recruitment to Ribosomes and Regulators of Translation.” *Annual Review of Biochemistry* 68 (1): 913–63. doi:10.1146/annurev.biochem.68.1.913.

Godfray, H. Charles J., John R. Beddington, Ian R. Crute, Lawrence Haddad, David Lawrence, James F. Muir, Jules Pretty, Sherman Robinson, Sandy M. Thomas, and Camilla Toulmin. 2010. “Food Security: The Challenge of Feeding 9 Billion People.” *Science* 327 (5967): 812–18. doi:10.1126/science.1185383.

Gourdon, Geneviève, François Radvanyi, Anne-Sophie Lia, Chantal Duros, Martine Blanche, Marc Abitbol, Claudine Junien, and Hlène Hofmann-Radvanyi. 1997. “Moderate Intergenerational and Somatic Instability of a 55-CTG Repeat in Transgenic Mice.” *Nature Genetics* 15 (2): 190–92. doi:10.1038/ng0297-190.

Granneman, Sander, and Susan J Baserga. 2004. “Ribosome Biogenesis: Of Knobs and RNA Processing.” *Experimental Cell Research* 296 (1): 43–50. doi:10.1016/j.yexcr.2004.03.016.

Grassucci, Robert A., Derek Taylor, and Joachim Frank. 2008. “Visualization of Macromolecular Complexes Using Cryo-Electron Microscopy with FEI Tecnai Transmission Electron Microscopes.” *Nature Protocols* 3 (2): 330–39. doi:10.1038/nprot.2007.474.

Greco, W. R., G. Bravo, and J. C. Parsons. 1995. “The Search for Synergy: A Critical Review from a Response Surface Perspective.” *Pharmacological Reviews* 47 (2): 331–85.

Gygi, Steven P., Beate Rist, Scott A. Gerber, Frantisek Turecek, Michael H. Gelb, and Ruedi Aebersold. 1999. “Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags.” *Nature Biotechnology* 17 (10): 994–99. doi:10.1038/13690.

Hallgrímsdóttir, Ingileif B., and Debbie S. Yuster. 2008. “A Complete Classification of Epistatic Two-Locus Models.” *BMC Genetics* 9 (1): 17. doi:10.1186/1471-2156-9-17.

Hanash, Samir M., Sharon J. Pitteri, and Vitor M. Faca. 2008. “Mining the Plasma Proteome for Cancer Biomarkers.” *Nature* 452 (7187): 571–79. doi:10.1038/nature06916.

Han, Hong, Manuel Irimia, P. Joel Ross, Hoon-Ki Sung, Babak Alipanahi, Laurent David, Azadeh Golipour, et al. 2013. “MBNL Proteins Repress ES-Cell-Specific Alternative Splicing and Reprogramming.” *Nature* 498 (7453): 241–45. doi:10.1038/nature12270.

Hanjra, Munir A., and M. Ejaz Qureshi. 2010. “Global Water Crisis and Future Food Security in an Era of Climate Change.” *Food Policy* 35 (5): 365–77. doi:10.1016/j.foodpol.2010.05.006.

Han, Xuemei, Aaron Aslanian, and John R Yates III. 2008. “Mass Spectrometry for Proteomics.” *Current Opinion in Chemical Biology*, Analytical Techniques/Mechanisms, 12 (5): 483–90. doi:10.1016/j.cbpa.2008.07.024.

Harris, S., C. Moncrieff, and K. Johnson. 1996. “Myotonic Dystrophy: Will the Real Gene Please Step Forward!” *Human Molecular Genetics* 5 Spec No: 1417–23.

Hazlehurst, Lori A., Terry H. Landowski, and William S. Dalton. 2003. “Role of the Tumor Microenvironment in Mediating de Novo Resistance to Drugs and Physiological Mediators of Cell Death.” *Oncogene* 22 (47): 7396–7402. doi:10.1038/sj.onc.1206943.

Hebert, Alexander S., Anna E. Merrill, Derek J. Bailey, Amelia J. Still, Michael S. Westphall, Eric R. Strieter, David J. Pagliarini, and Joshua J. Coon. 2013. “Neutron-Encoded Mass Signatures for Multiplexed Proteome Quantification.” *Nature Methods* 10 (4): 332–34. doi:10.1038/nmeth.2378.

Hennrich, Marco Leo, and Anne-Claude Gavin. 2015. “Quantitative Mass Spectrometry of Posttranslational Modifications: Keys to Confidence.” *Science Signaling* 8 (371): re5–re5. doi:10.1126/scisignal.aaa6466.

Hermens, Joop, Peter Leeuwangh, and Aalt Musch. 1985. “Joint Toxicity of Mixtures of Groups of Organic Aquatic Pollutants to the Guppy (Poecilia Reticulata).” *Ecotoxicology and Environmental Safety* 9 (3): 321–26. doi:10.1016/0147-6513(85)90049-1.

Hinnebusch, Alan G. 2005. “Translational Regulation of Gcn4 and the General Amino Acid Control of Yeast*.” *Annual Review of Microbiology* 59 (1): 407–50. doi:10.1146/annurev.micro.59.031805.133833.

———. 2015. “Translational Control 1995–2015: Unveiling Molecular Underpinnings and Roles in Human Biology.” *RNA* 21 (4): 636–39. doi:10.1261/rna.049957.115.

Hodge, WG, JP Whitcher, and W Satariano. 1994. “Risk Factors for Age-Related Cataracts.” *Epidemiologic Reviews* 17 (2): 336–46.

Hoek, Kristen L., Parimal Samir, Leigh M. Howard, Xinnan Niu, Nripesh Prasad, Allison Galassie, Qi Liu, et al. 2015. “A Cell-Based Systems Biology Assessment of Human Blood to Monitor Immune Responses after Influenza Vaccination.” *PLoS ONE* 10 (2): e0118528. doi:10.1371/journal.pone.0118528.

Hogquist, Kristin A., Stephen C. Jameson, William R. Heath, Jane L. Howard, Michael J. Bevan, and Francis R. Carbone. 1994. “T Cell Receptor Antagonist Peptides Induce Positive Selection.” *Cell* 76 (1): 17–27. doi:10.1016/0092-8674(94)90169-4.

Ho, Thai H., Nicolas Charlet-B, Michael G. Poulos, Gopal Singh, Maurice S. Swanson, and Thomas A. Cooper. 2004. “Muscleblind Proteins Regulate Alternative Splicing.” *The EMBO Journal* 23 (15): 3103–12. doi:10.1038/sj.emboj.7600300.

Hsiao, K.M., S.S. Chen, S.Y. Li, S.Y. Chiang, H.M. Lin, H. Pan, C.C. Huang, et al. 2003. “Epidemiological and Genetic Studies of Myotonic Dystrophy Type 1 in Taiwan.” *Neuroepidemiology* 22 (5): 283–89. doi:10.1159/000071191.

Hu, Michael C.-Y., Pedro Tranque, Gerald M. Edelman, and Vincent P. Mauro. 1999. “rRNA-Complementarity in the 5′ Untranslated Region of mRNA Specifying the Gtx Homeodomain Protein: Evidence That Base- Pairing to 18S rRNA Affects Translational Efficiency.” *Proceedings of the National Academy of Sciences* 96 (4): 1339–44. doi:10.1073/pnas.96.4.1339.

Hummel, Maureen, Jan H. G. Cordewener, Joost C. M. de Groot, Sjef Smeekens, Antoine H. P. America, and Johannes Hanson. 2012. “Dynamic Protein Composition of Arabidopsis Thaliana Cytosolic Ribosomes in Response to Sucrose Feeding as Revealed by Label Free MSE Proteomics.” *PROTEOMICS* 12 (7): 1024–38. doi:10.1002/pmic.201100413.

Hunt, D. F., J. R. Yates, J. Shabanowitz, S. Winston, and C. R. Hauer. 1986. “Protein Sequencing by Tandem Mass Spectrometry.” *Proceedings of the National Academy of Sciences* 83 (17): 6233–37.

Ideker, Trey, Timothy Galitski, and Leroy Hood. 2001. “A New Approach to Decoding Life: Systems Biology.” *Annual Review of Genomics and Human Genetics* 2 (1): 343–72. doi:10.1146/annurev.genom.2.1.343.

Inge-Vechtomov, Sergei, Galina Zhouravleva, and Michel Philippe. 2003. “Eukaryotic Release Factors (eRFs) History.” *Biology of the Cell*, Post-transcriptional control of gene expression in cell function, 95 (3–4): 195–209. doi:10.1016/S0248-4900(03)00035-2.

Ingolia, Nicholas T. 2014. “Ribosome Profiling: New Views of Translation, from Single Codons to Genome Scale.” *Nature Reviews Genetics* 15 (3): 205–13. doi:10.1038/nrg3645.

Jackson, Richard J., Christopher U. T. Hellen, and Tatyana V. Pestova. 2012. “Termination and Post-Termination Events in Eukaryotic Translation.” In *Advances in Protein Chemistry and Structural Biology*, edited by Assen Marintchev, 86:45–93. Fidelity and Quality Control in Gene Expression. Academic Press. http://www.sciencedirect.com/science/article/pii/B9780123864970000025.

Jang, S. K., H. G. Kräusslich, M. J. Nicklin, G. M. Duke, A. C. Palmenberg, and E. Wimmer. 1988. “A Segment of the 5’ Nontranslated Region of Encephalomyocarditis Virus RNA Directs Internal Entry of Ribosomes during in Vitro Translation.” *Journal of Virology* 62 (8): 2636–43.

Jansen, Gert, Patricia J. T. A. Groenen, Dietmar Bächner, Paul H. K. Jap, Marga Coerwinkel, Frank Oerlemans, Walther van den Broek, et al. 1996. “Abnormal Myotonic Dystrophy Protein Kinase Levels Produce Only Mild Myopathy in Mice.” *Nature Genetics* 13 (3): 316–24. doi:10.1038/ng0796-316.

Jansen, G., P. Willems, M. Coerwinkel, W. Nillesen, H. Smeets, L. Vits, C. Höweler, H. Brunner, and B. Wieringa. 1994. “Gonosomal Mosaicism in Myotonic Dystrophy Patients: Involvement of Mitotic Events in (CTG)n Repeat Variation and Selection against Extreme Expansion in Sperm.” *American Journal of Human Genetics* 54 (4): 575–85.

Johnson, Arlen W, Elsebet Lund, and James Dahlberg. 2002. “Nuclear Export of Ribosomal Subunits.” *Trends in Biochemical Sciences* 27 (11): 580–85. doi:10.1016/S0968-0004(02)02208-9.

Kaerlein, Margret, and Ivan Horak. 1976. “Phosphorylation of Ribosomal Proteins in HeLa Cells Infected with Vaccinia Virus.” *Nature* 259 (5539): 150–51. doi:10.1038/259150a0.

Kanadia, Rahul N, Carl R Urbinati, Valerie J Crusselle, Defang Luo, Young-Jae Lee, Jeffrey K Harrison, S. Paul Oh, and Maurice S Swanson. 2003. “Developmental Expression of Mouse Muscleblind Genes Mbnl1, Mbnl2 and Mbnl3.” *Gene Expression Patterns* 3 (4): 459–62. doi:10.1016/S1567-133X(03)00064-4.

Kanani, Harin, Bhaskar Dutta, and Maria I. Klapa. 2010. “Individual vs. Combinatorial Effect of Elevated CO2 Conditions and Salinity Stress on Arabidopsis Thaliana Liquid Cultures: Comparing the Early Molecular Response Using Time-Series Transcriptomic and Metabolomic Analyses.” *BMC Systems Biology* 4 (1): 177. doi:10.1186/1752-0509-4-177.

Kapasi, Purvi, Sujan Chaudhuri, Keyur Vyas, Diane Baus, Anton A. Komar, Paul L. Fox, William C. Merrick, and Barsanjit Mazumder. 2007. “L13a Blocks 48S Assembly: Role of a General Initiation Factor in mRNA-Specific Translational Control.” *Molecular Cell* 25 (1): 113–26. doi:10.1016/j.molcel.2006.11.028.

Kapp, Lee D, and Jon R Lorsch. 2004. “The Molecular Mechanics of Eukaryotic Translation.” *Annual Review of Biochemistry* 73 (1): 657–704.

Kenny, Paraic A., Genee Y. Lee, and Mina J. Bissell. 2007. “Targeting the Tumor Microenvironment.” *Frontiers in Bioscience : A Journal and Virtual Library* 12 (May): 3468–74.

Kessenbrock, Kai, Vicki Plaks, and Zena Werb. 2010. “Matrix Metalloproteinases: Regulators of the Tumor Microenvironment.” *Cell* 141 (1): 52–67. doi:10.1016/j.cell.2010.03.015.
Kim, Min-Sik, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, et al. 2014. “A Draft Map of the Human Proteome.” *Nature* 509 (7502): 575–81. doi:10.1038/nature13302.
Kitano, Hiroaki. 2000. “Perspectives on Systems Biology.” *New Generation Computing* 18 (3): 199–216. doi:10.1007/BF03037529.
———. 2002. “Systems Biology: A Brief Overview.” *Science* 295 (5560): 1662–64. doi:10.1126/science.1069492.
Klesert, Todd R., Diane H. Cho, John I. Clark, James Maylie, John Adelman, Lauren Snider, Eric C. Yuen, Philippe Soriano, and Stephen J. Tapscott. 2000. “Mice Deficient in Six5 Develop Cataracts: Implications for Myotonic Dystrophy.” *Nature Genetics* 25 (1): 105–9. doi:10.1038/75490.
Klesert, Todd R., Anne D. Otten, Thomas D. Bird, and Stephen J. Tapscott. 1997. “Trinucleotide Repeat Expansion at the Myotonic Dystrophy Locus Reduces Expression of DMAHP.” *Nature Genetics* 16 (4): 402–6. doi:10.1038/ng0897-402.
Knijnenburg, Theo A., Jean-Marc G. Daran, Marcel A. van den Broek, Pascale AS Daran-Lapujade, Johannes H. de Winde, Jack T. Pronk, Marcel JT Reinders, and Lodewyk FA Wessels. 2009. “Combinatorial Effects of Environmental Parameters on Transcriptional Regulation in Saccharomyces Cerevisiae: A Quantitative Analysis of a Compendium of Chemostat-Based Transcriptome Data.” *BMC Genomics* 10 (1): 53. doi:10.1186/1471-2164-10-53.
Knijnenburg, Theo A., Johannes H. de Winde, Jean-Marc Daran, Pascale Daran-Lapujade, Jack T. Pronk, Marcel JT Reinders, and Lodewyk FA Wessels. 2007. “Exploiting Combinatorial Cultivation Conditions to Infer Transcriptional Regulation.” *BMC Genomics* 8 (1): 25. doi:10.1186/1471-2164-8-25.
Koch, M. C., K. Steinmeyer, C. Lorenz, K. Ricker, F. Wolf, M. Otto, B. Zoll, F. Lehmann-Horn, K. H. Grzeschik, and T. J. Jentsch. 1992. “The Skeletal Muscle Chloride Channel in Dominant and Recessive Human Myotonia.” *Science* 257 (5071): 797–800. doi:10.1126/science.1379744.
Kolaczyk, Eric D., and Gábor Csárdi. 2014. *Statistical Analysis of Network Data with R*. Vol. 65. Use R! New York, NY: Springer New York. http://link.springer.com/10.1007/978-1-4939-0983-4.
Kolkman, Annemieke, Pascale Daran-Lapujade, Asier Fullaondo, Maurien M A Olsthoorn, Jack T Pronk, Monique Slijper, and Albert J R Heck. 2006. “Proteome Analysis of Yeast Response to Various Nutrient Limitations.” *Molecular Systems Biology* 2 (1): n/a – n/a. doi:10.1038/msb4100069.
Komili, Suzanne, Natalie G. Farny, Frederick P. Roth, and Pamela A. Silver. 2007. “Functional Specificity among Ribosomal Proteins Regulates Gene Expression.” *Cell* 131 (3): 557–71. doi:10.1016/j.cell.2007.08.037.
Kondrashov, Nadya, Aya Pusic, Craig R. Stumpf, Kunihiko Shimizu, Andrew C. Hsieh, Shifeng Xue, Junko Ishijima, Toshihiko Shiroishi, and Maria Barna. 2011. “Ribosome-Mediated Specificity in Hox mRNA Translation and Vertebrate Tissue Patterning.” *Cell* 145 (3): 383–97. doi:10.1016/j.cell.2011.03.028.
Korneluk, R. G., A. E. MacKenzie, Y. Nakamura, I. Dubé, P. Jacob, and A. G. W. Hunter. 1989. “A Reordering of Human Chromosome 19 Long-Arm DNA Markers and Identification of Markers Flanking the Myotonic Dystrophy Locus.” *Genomics* 5 (3): 596–604. doi:10.1016/0888-7543(89)90028-1.
Korostelev, Andrei A. 2014. “A Deeper Look into Translation Initiation.” *Cell* 159 (3): 475–76. doi:10.1016/j.cell.2014.10.005.

Kostrzewa, Markus, Uta Burck-Lehmann, and Ulrich Müller. 1994. “Autosomal Dominant Amyotrophic Lateral Sclerosis: A Novel Mutation in the Cu/Zn Superoxide Dismutase-1 Gene.” *Human Molecular Genetics* 3 (12): 2261–62. doi:10.1093/hmg/3.12.2261.
Kressler, Dieter, Patrick Linder, and Jesús de la Cruz. 1999. “Protein Trans-Acting Factors Involved in Ribosome Biogenesis in Saccharomyces Cerevisiae.” *Molecular and Cellular Biology* 19 (12): 7897–7912. doi:10.1128/MCB.19.12.7897.
Kruiswijk, T, A Kunst, R J Planta, and W H Mager. 1978. “Modification of Yeast Ribosomal Proteins. Methylation.” *Biochemical Journal* 175 (1): 221–25.
Krzywinski, Martin I., Jacqueline E. Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. “Circos: An Information Aesthetic for Comparative Genomics.” *Genome Research*, June. doi:10.1101/gr.092759.109.
Kurland, C. G., P. Voynow, S. J. S. Hardy, L. Randall, and L. Lutter. 1969. “Physical and Functional Heterogeneity of E. Coli Ribosomes.” *Cold Spring Harbor Symposia on Quantitative Biology* 34 (January): 17–24. doi:10.1101/SQB.1969.034.01.006.
Kurosaki, Tatsuaki, Shintaroh Ueda, Takafumi Ishida, Koji Abe, Kinji Ohno, and Tohru Matsuura. 2012. “The Unstable CCTG Repeat Responsible for Myotonic Dystrophy Type 2 Originates from an AluSx Element Insertion into an Early Primate Genome.” *PLoS ONE* 7 (6): e38379. doi:10.1371/journal.pone.0038379.
Ladd, Andrea N., Nicolas Charlet-B, and Thomas A. Cooper. 2001. “The CELF Family of RNA Binding Proteins Is Implicated in Cell-Specific and Developmentally Regulated Alternative Splicing.” *Molecular and Cellular Biology* 21 (4): 1285–96. doi:10.1128/MCB.21.4.1285-1296.2001.
Lam, Sik Lok, Feng Wu, Hao Yang, and Lai Man Chi. 2011. “The Origin of Genetic Instability in CCTG Repeats.” *Nucleic Acids Research* 39 (14): 6260–68. doi:10.1093/nar/gkr185.
Larance, Mark, and Angus I. Lamond. 2015. “Multidimensional Proteomics for Cell Biology.” *Nature Reviews Molecular Cell Biology* 16 (5): 269–80. doi:10.1038/nrm3970.
Lavedan, C, H Hofmann-Radvanyi, P Shelbourne, J P Rabes, C Duros, D Savoy, I Dehaupas, S Luce, K Johnson, and C Junien. 1993. “Myotonic Dystrophy: Size- and Sex-Dependent Dynamics of CTG Meiotic Instability, and Somatic Mosaicism.” *American Journal of Human Genetics* 52 (5): 875–83.
Law, Brian K. 2005. “Rapamycin: An Anti-Cancer Immunosuppressant?” *Critical Reviews in Oncology/Hematology*, Immunosuppressive Treatment and Induction of Cancer, 56 (1): 47–60. doi:10.1016/j.critrevonc.2004.09.009.
Lee, Amy Si-Ying, Rebeca Burdeinick-Kerr, and Sean P. J. Whelan. 2013. “A Ribosome-Specialized Translation Initiation Pathway Is Required for Cap-Dependent Translation of Vesicular Stomatitis Virus mRNAs.” *Proceedings of the National Academy of Sciences* 110 (1): 324–29. doi:10.1073/pnas.1216454109.
Lee, Johanna E., and Thomas A. Cooper. 2009. “Pathogenic Mechanisms of Myotonic Dystrophy.” *Biochemical Society Transactions* 37 (0 6). doi:10.1042/BST0371281.
Lee, Jong-Hyeon, and Peter F. Landrum. 2006. “Development of a Multi-Component Damage Assessment Model (MDAM) for Time-Dependent Mixture Toxicity with Toxicokinetic Interactions.” *Environmental Science & Technology* 40 (4): 1341–49. doi:10.1021/es051120f.
Ligges, Uwe, and Martin Mächler. 2003. “Scatterplot3d – an R Package for Visualizing Multivariate Data.” *Journal of Statistical Software* 8 (11): 1–20.
Lindquist, S. 1986. “The Heat-Shock Response.” *Annual Review of Biochemistry* 55 (1): 1151–91. doi:10.1146/annurev.bi.55.070186.005443.
Link, Andrew J., Jimmy Eng, David M. Schieltz, Edwin Carmack, Gregory J. Mize, David R. Morris, Barbara M. Garvik, and John R. Yates. 1999. “Direct Analysis of Protein

Complexes Using Mass Spectrometry." *Nature Biotechnology* 17 (7): 676–82. doi:10.1038/10890.
Liquori, Christina L., Yoshio Ikeda, Marcy Weatherspoon, Kenneth Ricker, Benedikt G. H. Schoser, Joline C. Dalton, John W. Day, and Laura P. W. Ranum. 2003. "Myotonic Dystrophy Type 2: Human Founder Haplotype and Evolutionary Conservation of the Repeat Tract." *The American Journal of Human Genetics* 73 (4): 849–62. doi:10.1086/378720.
Liquori, Christina L., Kenneth Ricker, Melinda L. Moseley, Jennifer F. Jacobsen, Wolfram Kress, Susan L. Naylor, John W. Day, and Laura P. W. Ranum. 2001. "Myotonic Dystrophy Type 2 Caused by a CCTG Expansion in Intron 1 of ZNF9." *Science* 293 (5531): 864–67. doi:10.1126/science.1062125.
Lister, Ryan, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. 2008. "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis." *Cell* 133 (3): 523–36. doi:10.1016/j.cell.2008.03.029.
Li, Wentian, and Jens Reich. 2000. "A Complete Enumeration and Classification of Two-Locus Disease Models." *Human Heredity* 50 (6): 334–49. doi:10.1159/000022939.
Lockhart, D. J., H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, et al. 1996. "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays." *Nature Biotechnology* 14 (13): 1675–80. doi:10.1038/nbt1296-1675.
Lodish, Harvey F. 1981. "Post-Translational Modification of Proteins." *Enzyme and Microbial Technology* 3 (3): 178–88. doi:10.1016/0141-0229(81)90084-3.
Loro, E., F. Rinaldi, A. Malena, E. Masiero, G. Novelli, C. Angelini, V. Romeo, M. Sandri, A. Botta, and L. Vergani. 2010. "Normal Myogenesis and Increased Apoptosis in Myotonic Dystrophy Type-1 Muscle Cells." *Cell Death & Differentiation* 17 (8): 1315–24. doi:10.1038/cdd.2010.33.
Machuca-Tzili, Laura, David Brook, and David Hilton-Jones. 2005. "Clinical and Molecular Aspects of the Myotonic Dystrophies: A Review." *Muscle & Nerve* 32 (1): 1–18. doi:10.1002/mus.20301.
MacLean, Brendan, Daniela M. Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L. Finney, Barbara Frewen, Randall Kern, David L. Tabb, Daniel C. Liebler, and Michael J. MacCoss. 2010. "Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments." *Bioinformatics* 26 (7): 966–68. doi:10.1093/bioinformatics/btq054.
Mahadevan, M., C. Tsilfidis, L. Sabourin, G. Shutler, C. Amemiya, G. Jansen, C. Neville, M. Narang, J. Barceló, and K. O'Hoy. 1992. "Myotonic Dystrophy Mutation: An Unstable CTG Repeat in the 3' Untranslated Region of the Gene." *Science (New York, N.Y.)* 255 (5049): 1253–55.
Makarova, Olga V, Evgeny M Makarov, Henning Urlaub, Cindy L Will, Marc Gentzel, Matthias Wilm, and Reinhard Lührmann. 2004. "A Subset of Human 35S U5 Proteins, Including Prp19, Function prior to Catalytic Step 1 of Splicing." *The EMBO Journal* 23 (12): 2381–91. doi:10.1038/sj.emboj.7600241.
Mani, Ramamurthy, Robert P. St.Onge, John L. Hartman, Guri Giaever, and Frederick P. Roth. 2008. "Defining Genetic Interaction." *Proceedings of the National Academy of Sciences* 105 (9): 3461–66. doi:10.1073/pnas.0712255105.
Mankodi, Ami, Eric Logigian, Linda Callahan, Carolyn McClain, Robert White, Don Henderson, Matt Krym, and Charles A. Thornton. 2000. "Myotonic Dystrophy in Transgenic Mice Expressing an Expanded CUG Repeat." *Science* 289 (5485): 1769–72. doi:10.1126/science.289.5485.1769.
Mankodi, Ami, Masanori P. Takahashi, Hong Jiang, Carol L. Beck, William J. Bowers, Richard T. Moxley, Stephen C. Cannon, and Charles A. Thornton. 2002. "Expanded CUG

Repeats Trigger Aberrant Splicing of ClC-1 Chloride Channel Pre-mRNA and Hyperexcitability of Skeletal Muscle in Myotonic Dystrophy." *Molecular Cell* 10 (1): 35–44. doi:10.1016/S1097-2765(02)00563-4.
Mankodi, Ami, Patana Teng-Umnuay, Matt Krym, Don Henderson, Maurice Swanson, and Charles A. Thornton. 2003. "Ribonuclear Inclusions in Skeletal Muscle in Myotonic Dystrophy Types 1 and 2." *Annals of Neurology* 54 (6): 760–68. doi:10.1002/ana.10763.
Martin, Ian, Jungwoo Wren Kim, Byoung Dae Lee, Ho Chul Kang, Jin-Chong Xu, Hao Jia, Jeannette Stankowski, et al. 2014. "Ribosomal Protein s15 Phosphorylation Mediates LRRK2 Neurodegeneration in Parkinson's Disease." *Cell* 157 (2): 472–85. doi:10.1016/j.cell.2014.01.064.
Martorell, L., J. M. Martinez, N. Carey, K. Johnson, and M. Baiget. 1995. "Comparison of CTG Repeat Length Expansion and Clinical Progression of Myotonic Dystrophy over a Five Year Period." *Journal of Medical Genetics* 32 (8): 593–96. doi:10.1136/jmg.32.8.593.
Martorell, Loreto, Keith Johnson, Catherine A. Boucher, and Montserrat Baiget. 1997. "Somatic Instability of the Myotonic Dystrophy (CTG)n Repeat during Human Fetal Development." *Human Molecular Genetics* 6 (6): 877–80. doi:10.1093/hmg/6.6.877.
Masuda, Akio, Henriette Skovgaard Andersen, Thomas Koed Doktor, Takaaki Okamoto, Mikako Ito, Brage Storstein Andresen, and Kinji Ohno. 2012. "CUGBP1 and MBNL1 Preferentially Bind to 3′ UTRs and Facilitate mRNA Decay." *Scientific Reports* 2 (January). doi:10.1038/srep00209.
Matsumura, Yoshihiro, Juro Sakai, and William R. Skach. 2013. "Endoplasmic Reticulum Protein Quality Control Is Determined by Cooperative Interactions between Hsp/c70 Protein and the CHIP E3 Ligase." *Journal of Biological Chemistry* 288 (43): 31069–79. doi:10.1074/jbc.M113.479345.
Mauro, Vincent P., and Gerald M. Edelman. 1997. "rRNA-like Sequences Occur in Diverse Primary Transcripts: Implications for the Control of Gene Expression." *Proceedings of the National Academy of Sciences* 94 (2): 422–27.
———. 2002. "The Ribosome Filter Hypothesis." *Proceedings of the National Academy of Sciences* 99 (19): 12031–36. doi:10.1073/pnas.192442499.
Mazumder, Barsanjit, Prabha Sampath, Vasudevan Seshadri, Ratan K Maitra, Paul E DiCorleto, and Paul L Fox. 2003. "Regulated Release of L13a from the 60S Ribosomal Subunit as A Mechanism of Transcript-Specific Translational Control." *Cell* 115 (2): 187–98. doi:10.1016/S0092-8674(03)00773-6.
Mbeunkui, Flaubert, and Donald J. Johann Jr. 2008. "Cancer and the Tumor Microenvironment: A Review of an Essential Relationship." *Cancer Chemotherapy and Pharmacology* 63 (4): 571–82. doi:10.1007/s00280-008-0881-9.
McIntosh, Kerri B., and Jonathan R. Warner. 2007. "Yeast Ribosomes: Variety Is the Spice of Life." *Cell* 131 (3): 450–51. doi:10.1016/j.cell.2007.10.028.
Meola, Giovanni, Valeria Sansone, Stefania Radice, Shana Skradski, and Louis Ptacek. 1996. "A Family with an Unusual Myotonic and Myopathic Phenotype and No CTG Expansion (proximal Myotonic Myopathy Syndrome): A Challenge for Future Molecular Studies." *Neuromuscular Disorders* 6 (3): 143–50. doi:10.1016/0960-8966(95)00040-2.
Merrick, William C. 2004. "Cap-Dependent and Cap-Independent Translation in Eukaryotic Systems." *Gene* 332 (May): 1–11. doi:10.1016/j.gene.2004.02.051.
Michalowski, Susan, Jill W. Miller, Carl R. Urbinati, Miltiadis Paliouras, Maurice S. Swanson, and Jack Griffith. 1999. "Visualization of Double-Stranded RNAs from the Myotonic Dystrophy Protein Kinase Gene and Interactions with CUG-Binding Protein." *Nucleic Acids Research* 27 (17): 3534–42. doi:10.1093/nar/27.17.3534.
Michalski, Annette, Eugen Damoc, Jan-Peter Hauschild, Oliver Lange, Andreas Wieghaus, Alexander Makarov, Nagarjuna Nagaraj, Juergen Cox, Matthias Mann, and Stevan Horning. 2011. "Mass Spectrometry-Based Proteomics Using Q Exactive, a High-

Performance Benchtop Quadrupole Orbitrap Mass Spectrometer." *Molecular & Cellular Proteomics* 10 (9): M111.011015. doi:10.1074/mcp.M111.011015.
Miller, J. W. 2000. "Recruitment of Human Muscleblind Proteins to (CUG)n Expansions Associated with Myotonic Dystrophy." *The EMBO Journal* 19 (17): 4439–48. doi:10.1093/emboj/19.17.4439.
Mizoguchi, Tsuyoshi, Kazuya Ichimura, and Kazuo Shinozaki. 1997. "Environmental Stress Response in Plants: The Role of Mitogen-Activated Protein Kinases." *Trends in Biotechnology* 15 (1): 15–19. doi:10.1016/S0167-7799(96)10074-3.
Monckton, Darren G., Mary I. Coolbaugh, Ken T. Ashizawa, Michael J. Siciliano, and C. Thomas Caskey. 1997. "Hypermutable Myotonic Dystrophy CTG Repeats in Transgenic Mice." *Nature Genetics* 15 (2): 193–96. doi:10.1038/ng0297-193.
Montojo, J., K. Zuberi, H. Rodriguez, F. Kazi, G. Wright, S. L. Donaldson, Q. Morris, and G. D. Bader. 2010. "GeneMANIA Cytoscape Plugin: Fast Gene Function Predictions on the Desktop." *Bioinformatics* 26 (22): 2927–28. doi:10.1093/bioinformatics/btq562.
Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. doi:10.1038/nmeth.1226.
Mostafavi, Sara, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. 2008. "GeneMANIA: A Real-Time Multiple Association Network Integration Algorithm for Predicting Gene Function." *Genome Biology* 9 (Suppl 1): S4. doi:10.1186/gb-2008-9-s1-s4.
Moxley III, Richard T. 1996. "Proximal Myotonic Myopathy: Mini-Review of a Recently Delineated Clinical Disorder." *Neuromuscular Disorders*, Symposium on Recent Advances in Diagnosis and Therapy of Neuromuscular Diseases, 6 (2): 87–93. doi:10.1016/0960-8966(95)00036-4.
Mukhopadhyay, Debnath, Courtney W. Houchen, Susan Kennedy, Brian K. Dieckgraefe, and Shrikant Anant. 2003. "Coupled mRNA Stabilization and Translational Silencing of Cyclooxygenase-2 by a Novel RNA Binding Protein, CUGBP2." *Molecular Cell* 11 (1): 113–26. doi:10.1016/S1097-2765(03)00012-1.
Mukhopadhyay, Rupak, Jie Jia, Abul Arif, Partho Sarothi Ray, and Paul L. Fox. 2009. "The GAIT System: A Gatekeeper of Inflammatory Gene Expression." *Trends in Biochemical Sciences* 34 (7): 324–31. doi:10.1016/j.tibs.2009.03.004.
Murray, John Isaac, Michael L. Whitfield, Nathan D. Trinklein, Richard M. Myers, Patrick O. Brown, and David Botstein. 2004. "Diverse and Specific Gene Expression Responses to Stresses in Cultured Human Cells." *Molecular Biology of the Cell* 15 (5): 2361–74. doi:10.1091/mbc.E03-11-0799.
Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. 2008. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." *Science* 320 (5881): 1344–49. doi:10.1126/science.1158441.
Nagano, Atsushi J., Yutaka Sato, Motohiro Mihara, Baltazar A. Antonio, Ritsuko Motoyama, Hironori Itoh, Yoshiaki Nagamura, and Takeshi Izawa. 2012. "Deciphering and Prediction of Transcriptome Dynamics under Fluctuating Field Conditions." *Cell* 151 (6): 1358–69. doi:10.1016/j.cell.2012.10.048.
Nakajima, J., N. Okamoto, J. Tohyama, M. Kato, H. Arai, O. Funahashi, Y. Tsurusaki, et al. 2015. "De Novo EEF1A2 Mutations in Patients with Characteristic Facial Features, Intellectual Disability, Autistic Behaviors and Epilepsy." *Clinical Genetics* 87 (4): 356–61. doi:10.1111/cge.12394.
Nakano, R., S. Sato, T. Inuzuka, K. Sakimura, M. Mishina, H. Takahashi, F. Ikuta, et al. 1994. "A Novel Mutation in Cu/Zn Superoxide Dismutase Gene in Japanese Familial

Amyotrophic Lateral Sclerosis." *Biochemical and Biophysical Research Communications* 200 (2): 695–703. doi:10.1006/bbrc.1994.1506.
Nakao, Akihiro, Maki Yoshihama, and Naoya Kenmochi. 2004. "RPG: The Ribosomal Protein Gene Database." *Nucleic Acids Research* 32 (suppl 1): D168–70. doi:10.1093/nar/gkh004.
Narla, Anupama, and Benjamin L. Ebert. 2010. "Ribosomopathies: Human Disorders of Ribosome Dysfunction." *Blood* 115 (16): 3196–3205. doi:10.1182/blood-2009-10-178129.
Nazar, Ross. 2004. "Ribosomal RNA Processing and Ribosome Biogenesis in Eukaryotes." *IUBMB Life* 56 (8): 457–65. doi:10.1080/15216540400010867.
Neilson, Karlie A., Naveid A. Ali, Sridevi Muralidharan, Mehdi Mirzaei, Michael Mariani, Gariné Assadourian, Albert Lee, Steven C. van Sluyter, and Paul A. Haynes. 2011. "Less Label, More Free: Approaches in Label-Free Quantitative Mass Spectrometry." *PROTEOMICS* 11 (4): 535–53. doi:10.1002/pmic.201000553.
Nesterchuk, M.V., P.V. Sergiev, and O.A. Dontsova. 2011. "Posttranslational Modifications of Ribosomal Proteins in Escherichia Coli." *Acta Naturae* 3 (2): 22–33.
Nevoigt, Elke, and Ulf Stahl. 1997. "Osmoregulation and Glycerol Metabolism in the Yeast Saccharomyces Cerevisiae." *FEMS Microbiology Reviews* 21 (3): 231–41. doi:10.1016/S0168-6445(97)00058-2.
Nicola, Raffaele De, Lucie A. Hazelwood, Erik A. F. De Hulster, Michael C. Walsh, Theo A. Knijnenburg, Marcel J. T. Reinders, Graeme M. Walker, Jack T. Pronk, Jean-Marc Daran, and Pascale Daran-Lapujade. 2007. "Physiological and Transcriptional Responses of Saccharomyces Cerevisiae to Zinc Limitation in Chemostat Cultures." *Applied and Environmental Microbiology* 73 (23): 7680–92. doi:10.1128/AEM.01445-07.
Nieminen, Taina T., Marie-Françoise O'Donohue, Yunpeng Wu, Hannes Lohi, Stephen W. Scherer, Andrew D. Paterson, Pekka Ellonen, et al. 2014. "Germline Mutation of RPS20, Encoding a Ribosomal Protein, Causes Predisposition to Hereditary Nonpolyposis Colorectal Carcinoma Without DNA Mismatch Repair Deficiency." *Gastroenterology* 147 (3): 595–98.e5. doi:10.1053/j.gastro.2014.06.009.
Nilsson, Jakob, and Poul Nissen. 2005. "Elongation Factors on the Ribosome." *Current Opinion in Structural Biology*, Sequences and topology/Nucleic acids, 15 (3): 349–54. doi:10.1016/j.sbi.2005.05.004.
Nishikura, Kazuko. 2010. "Functions and Regulation of RNA Editing by ADAR Deaminases." *Annual Review of Biochemistry* 79 (1): 321–49. doi:10.1146/annurev-biochem-060208-105251.
Nürenberg, Elina, and Robert Tampé. 2013. "Tying up Loose Ends: Ribosome Recycling in Eukaryotes and Archaea." *Trends in Biochemical Sciences* 38 (2): 64–74. doi:10.1016/j.tibs.2012.11.003.
Nussinov, Ruth, Chung-Jung Tsai, Fuxiao Xin, and Predrag Radivojac. 2012. "Allosteric Post-Translational Modification Codes." *Trends in Biochemical Sciences* 37 (10): 447–55. doi:10.1016/j.tibs.2012.07.001.
Nyborg, Jens, and Anders Liljas. 1998. "Protein Biosynthesis: Structural Studies of the Elongation Cycle." *FEBS Letters* 430 (1–2): 95–99. doi:10.1016/S0014-5793(98)00624-3.
Odermatt, A., K. Barton, V. K. Khanna, J. Mathieu, D. Escolar, T. Kuntzer, G. Karpati, and D. H. MacLennan. 2000. "The Mutation of Pro789 to Leu Reduces the Activity of the Fast-Twitch Skeletal Muscle Sarco(endo)plasmic Reticulum Ca2+ ATPase (SERCA1) and Is Associated with Brody Disease." *Human Genetics* 106 (5): 482–91.
Ohn, Takbum, Nancy Kedersha, Tyler Hickman, Sarah Tisdale, and Paul Anderson. 2008. "A Functional RNAi Screen Links O-GlcNAc Modification of Ribosomal Proteins to Stress

Granule and Processing Body Assembly." *Nature Cell Biology* 10 (10): 1224–31. doi:10.1038/ncb1783.
Ohta, Akio, and Michail Sitkovsky. 2001. "Role of G-Protein-Coupled Adenosine Receptors in Downregulation of Inflammation and Protection from Tissue Damage." *Nature* 414 (6866): 916–20. doi:10.1038/414916a.
Ong, Shao-En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. 2002. "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics." *Molecular & Cellular Proteomics* 1 (5): 376–86. doi:10.1074/mcp.M200025-MCP200.
Ong, Shao-En, and Matthias Mann. 2005. "Mass Spectrometry–based Proteomics Turns Quantitative." *Nature Chemical Biology* 1 (5): 252–62. doi:10.1038/nchembio736.
Osborne, Robert J., Xiaoyan Lin, Stephen Welle, Krzysztof Sobczak, Jason R. O'Rourke, Maurice S. Swanson, and Charles A. Thornton. 2009. "Transcriptional and Post-Transcriptional Impact of Toxic RNA in Myotonic Dystrophy." *Human Molecular Genetics* 18 (8): 1471–81. doi:10.1093/hmg/ddp058.
Osborne, Robert J., and Charles A. Thornton. 2006. "RNA-Dominant Diseases." *Human Molecular Genetics* 15 (suppl 2): R162–69. doi:10.1093/hmg/ddl181.
Otten, A. D., and S. J. Tapscott. 1995. "Triplet Repeat Expansion in Myotonic Dystrophy Alters the Adjacent Chromatin Structure." *Proceedings of the National Academy of Sciences of the United States of America* 92 (12): 5465–69.
Owens, Geoffrey C., Stephen A. Chappell, Vincent P. Mauro, and Gerald M. Edelman. 2001. "Identification of Two Short Internal Ribosome Entry Sites Selected from Libraries of Random Oligonucleotides." *Proceedings of the National Academy of Sciences* 98 (4): 1471–76. doi:10.1073/pnas.98.4.1471.
Paillard, L. 1998. "EDEN and EDEN-BP, a Cis Element and an Associated Factor That Mediate Sequence-Specific mRNA Deadenylation in Xenopus Embryos." *The EMBO Journal* 17 (1): 278–87. doi:10.1093/emboj/17.1.278.
Pearson, Christopher E., Kerrie Nichol Edamura, and John D. Cleary. 2005. "Repeat Instability: Mechanisms of Dynamic Mutations." *Nature Reviews Genetics* 6 (10): 729–42. doi:10.1038/nrg1689.
Pelletier, Jerry, and Nahum Sonenberg. 1988. "Internal Initiation of Translation of Eukaryotic mRNA Directed by a Sequence Derived from Poliovirus RNA." *Nature* 334 (6180): 320–25. doi:10.1038/334320a0.
Peng, Jie, Pei Wang, Nengfeng Zhou, and Ji Zhu. 2009. "Partial Correlation Estimation by Joint Sparse Regression Models." *Journal of the American Statistical Association* 104 (486): 735–46. doi:10.1198/jasa.2009.0126.
Peterson, Thomas A., Asa Adadey, Ivette Santana-Cruz, Yanan Sun, Andrew Winder, and Maricel G. Kann. 2010. "DMDM: Domain Mapping of Disease Mutations." *Bioinformatics* 26 (19): 2458–59. doi:10.1093/bioinformatics/btq447.
Phillips, Patrick C. 1998. "The Language of Gene Interaction." *Genetics* 149 (3): 1167–71.
———. 2008. "Epistasis — the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems." *Nature Reviews Genetics* 9 (11): 855–67. doi:10.1038/nrg2452.
Phillips, Theresa. 2008. "The Role of Methylation in Gene Expression." *Nature Education* 1 (1): 116.
Picotti, Paola, Bernd Bodenmiller, Lukas N. Mueller, Bruno Domon, and Ruedi Aebersold. 2009. "Full Dynamic Range Proteome Analysis of S. Cerevisiae by Targeted Proteomics." *Cell* 138 (4): 795–806. doi:10.1016/j.cell.2009.05.051.
Pierce, Benjamin. 2005. *Genetics - A Conceptual Approach - 2nd (Second) Edition*.
Planta, Rudi J. 1997. "Regulation of Ribosome Synthesis in Yeast." *Yeast* 13 (16): 1505–18. doi:10.1002/(SICI)1097-0061(199712)13:16<1505::AID-YEA229>3.0.CO;2-I.

Pratt, Julie M., June Petty, Isabel Riba-Garcia, Duncan H. L. Robertson, Simon J. Gaskell, Stephen G. Oliver, and Robert J. Beynon. 2002. “Dynamics of Protein Turnover, a Missing Dimension in Proteomics.” *Molecular & Cellular Proteomics* 1 (8): 579–91. doi:10.1074/mcp.M200046-MCP200.

Pu, Shuye, Jim Vlasblom, Andrew Emili, Jack Greenblatt, and Shoshana J. Wodak. 2007. “Identifying Functional Modules in the Physical Interactome of Saccharomyces Cerevisiae.” *PROTEOMICS* 7 (6): 944–60. doi:10.1002/pmic.200600636.

Pu, Shuye, Jessica Wong, Brian Turner, Emerson Cho, and Shoshana J. Wodak. 2009. “Up-to-Date Catalogues of Yeast Protein Complexes.” *Nucleic Acids Research* 37 (3): 825–31. doi:10.1093/nar/gkn1005.

Ramagopal, S. 1991. “Covalent Modifications of Ribosomal Proteins in Growing and Aggregation-Competent Dictyostelium Discoideum: Phosphorylation and Methylation.” *Biochemistry and Cell Biology* 69 (4): 263–68. doi:10.1139/o91-040.

———. 1992. “Are Eukaryotic Ribosomes Heterogeneous? Affirmations on the Horizon.” *Biochemistry and Cell Biology* 70 (5): 269–72. doi:10.1139/o92-042.

Ramagopal, Subbanaidu, and Herbert L. Ennis. 1981. “Regulation of Synthesis of Cell-Specific Ribosomal Proteins during Differentiation of Dictyostelium Discoideum*.” *Proceedings of the National Academy of Sciences of the United States of America* 78 (5): 3083–87.

———. 1984. “Conservation and Variation of Ribosomal Proteins in Several Species of the Cellular Slime Molds Dictyostelium and Polysphondylium.” *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 805 (3): 300–305. doi:10.1016/0167-4889(84)90086-7.

Ranum, Laura P. W., and Thomas A. Cooper. 2006. “Rna-Mediated Neuromuscular Disorders.” *Annual Review of Neuroscience* 29 (1): 259–77. doi:10.1146/annurev.neuro.29.051605.113014.

Ranum, Laura P. W., and John W. Day. 2002. “Myotonic Dystrophy: Clinical and Molecular Parallels between Myotonic Dystrophy Type 1 and Type 2.” *Current Neurology and Neuroscience Reports* 2 (5): 465–70. doi:10.1007/s11910-002-0074-6.

Ranum, Laura P. W., Paul F. Rasmussen, Kellie A. Benzow, Michael D. Koob, and John W. Day. 1998. “Genetic Mapping of a Second Myotonic Dystrophy Locus.” *Nature Genetics* 19 (2): 196–98. doi:10.1038/570.

Rau, Frédérique, Fernande Freyermuth, Charlotte Fugier, Jean-Philippe Villemin, Marie-Christine Fischer, Bernard Jost, Doulaye Dembele, et al. 2011. “Misregulation of miR-1 Processing Is Associated with Heart Defects in Myotonic Dystrophy.” *Nature Structural & Molecular Biology* 18 (7): 840–45. doi:10.1038/nsmb.2067.

Ray, Partho Sarothi, Abul Arif, and Paul L. Fox. 2007. “Macromolecular Complexes as Depots for Releasable Regulatory Proteins.” *Trends in Biochemical Sciences* 32 (4): 158–64. doi:10.1016/j.tibs.2007.02.003.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Reddy, Sita, Daniel B. J. Smith, Mark M. Rich, John M. Leferovich, Patricia Reilly, Brigid M. Davis, Khoa Tran, et al. 1996. “Mice Lacking the Myotonic Dystrophy Protein Kinase Develop a Late Onset Progressive Myopathy.” *Nature Genetics* 13 (3): 325–35. doi:10.1038/ng0796-325.

Redman JB, Fenwick RG, Jr, Fu Y, Pizzuti A, and Caskey C. 1993. “Relationship between Parental Trinucleotide Gct Repeat Length and Severity of Myotonic Dystrophy in Offspring.” *JAMA* 269 (15): 1960–65. doi:10.1001/jama.1993.03500150072029.

Regenberg, Birgitte, Thomas Grotkjær, Ole Winther, Anders Fausbøll, Mats Åkesson, Christoffer Bro, Lars Kai Hansen, Søren Brunak, and Jens Nielsen. 2006. “Growth-Rate Regulated Genes Have Profound Impact on Interpretation of Transcriptome Profiling in

Saccharomyces Cerevisiae." *Genome Biology* 7 (11): 1–13. doi:10.1186/gb-2006-7-11-r107.
Rhodes, Jeremy D., Martin C. Lott, Sarah L. Russell, Vincent Moulton, Julie Sanderson, I. Michael Wormstone, and David C. Broadway. 2012. "Activation of the Innate Immune Response and Interferon Signalling in Myotonic Dystrophy Type 1 and Type 2 Cataracts." *Human Molecular Genetics* 21 (4): 852–62. doi:10.1093/hmg/ddr515.
Richter, Klaus, Martin Haslbeck, and Johannes Buchner. 2010. "The Heat Shock Response: Life on the Verge of Death." *Molecular Cell* 40 (2): 253–66. doi:10.1016/j.molcel.2010.10.006.
Ricker K, Koch MC, Lehmann-Horn F, and et al. 1995. "Proximal Myotonic Myopathy: Clinical Features of a Multisystem Disorder Similar to Myotonic Dystrophy." *Archives of Neurology* 52 (1): 25–31. doi:10.1001/archneur.1995.00540250029009.
Ricker, K., M. C. Koch, F. Lehmann-Horn, D. Pongratz, M. Otto, R. Heine, and R. T. Moxley. 1994. "Proximal Myotonic Myopathy A New Dominant Disorder with Myotonia, Muscle Weakness, and Cataracts." *Neurology* 44 (8): 1448–1448. doi:10.1212/WNL.44.8.1448.
Riezman, Howard. 2004. "Why Do Cells Require Heat Shock Proteins to Survive Heat Stress?" *Cell Cycle* 3 (1): 60–62. doi:10.4161/cc.3.1.625.
Robberecht, Wim, and Thomas Philips. 2013. "The Changing Scene of Amyotrophic Lateral Sclerosis." *Nature Reviews Neuroscience* 14 (4): 248–64. doi:10.1038/nrn3430.
Roberts, George G., and Alan P. Hudson. 2006. "Transcriptome Profiling of Saccharomyces Cerevisiae during a Transition from Fermentative to Glycerol-Based Respiratory Growth Reveals Extensive Metabolic and Structural Remodeling." *Molecular Genetics and Genomics* 276 (2): 170–86. doi:10.1007/s00438-006-0133-9.
Robertson, Keith D. 2005. "DNA Methylation and Human Disease." *Nature Reviews Genetics* 6 (8): 597–610. doi:10.1038/nrg1655.
Rosegrant, Mark W., and Sarah A. Cline. 2003. "Global Food Security: Challenges and Policies." *Science* 302 (5652): 1917–19. doi:10.1126/science.1092958.
Rosen, Daniel R., Teepu Siddique, David Patterson, Denise A. Figlewicz, Peter Sapp, Afif Hentati, Deirdre Donaldson, et al. 1993. "Mutations in Cu/Zn Superoxide Dismutase Gene Are Associated with Familial Amyotrophic Lateral Sclerosis." *Nature* 362 (6415): 59–62. doi:10.1038/362059a0.
Ross, Philip L., Yulin N. Huang, Jason N. Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, et al. 2004. "Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents." *Molecular & Cellular Proteomics* 3 (12): 1154–69. doi:10.1074/mcp.M400129-MCP200.
Roth, Udo, Edda von Roepenack-Lahaye, and Stephan Clemens. 2006. "Proteome Changes in Arabidopsis Thaliana Roots upon Exposure to Cd2+." *Journal of Experimental Botany* 57 (15): 4003–13. doi:10.1093/jxb/erl170.
Rouquette, Jacques, Valérie Choesmel, and Pierre-Emmanuel Gleizes. 2005. "Nuclear Export and Cytoplasmic Processing of Precursors to the 40S Ribosomal Subunits in Mammalian Cells." *The EMBO Journal* 24 (16): 2862–72. doi:10.1038/sj.emboj.7600752.
Rudnick, Paul A., Karl R. Clauser, Lisa E. Kilpatrick, Dmitrii V. Tchekhovskoi, Pedatsur Neta, Nikša Blonder, Dean D. Billheimer, et al. 2010. "Performance Metrics for Liquid Chromatography-Tandem Mass Spectrometry Systems in Proteomics Analyses." *Molecular & Cellular Proteomics* 9 (2): 225–41. doi:10.1074/mcp.M900223-MCP200.
Ruggero, Davide, and Pier Paolo Pandolfi. 2003. "Does the Ribosome Translate Cancer?" *Nature Reviews Cancer* 3 (3): 179–92. doi:10.1038/nrc1015.
Ruvinsky, Igor, and Oded Meyuhas. 2006. "Ribosomal Protein S6 Phosphorylation: From Protein Synthesis to Cell Size." *Trends in Biochemical Sciences* 31 (6): 342–48. doi:10.1016/j.tibs.2006.04.003.

Ruvinsky, Igor, Nitzan Sharon, Tal Lerer, Hannah Cohen, Miri Stolovich-Rain, Tomer Nir, Yuval Dor, Philip Zisman, and Oded Meyuhas. 2005. “Ribosomal Protein S6 Phosphorylation Is a Determinant of Cell Size and Glucose Homeostasis.” *Genes & Development* 19 (18): 2199–2211. doi:10.1101/gad.351605.

Saito, Tsukasa, Yoshinobu Amakusa, Takashi Kimura, Osamu Yahara, Hitoshi Aizawa, Yoshio Ikeda, John W. Day, Laura P. W. Ranum, Kinji Ohno, and Tohru Matsuura. 2007. “Myotonic Dystrophy Type 2 in Japan: Ancestral Origin Distinct from Caucasian Families.” *Neurogenetics* 9 (1): 61–63. doi:10.1007/s10048-007-0110-4.

Saldanha, Alok J. 2004. “Java Treeview—extensible Visualization of Microarray Data.” *Bioinformatics* 20 (17): 3246–48. doi:10.1093/bioinformatics/bth349.

Samir, Parimal, Rahul, James C. Slaughter, and Andrew J. Link. 2015. “Environmental Interactions and Epistasis Are Revealed in the Proteomic Responses to Complex Stimuli.” Edited by Ben Lehner. *PLOS ONE* 10 (8): e0134099. doi:10.1371/journal.pone.0134099.

Sammons, Morgan A., Amanda K. Antons, Mourad Bendjennat, Bjarne Udd, Ralf Krahe, and Andrew J. Link. 2010. “ZNF9 Activation of IRES-Mediated Translation of the Human ODC mRNA Is Decreased in Myotonic Dystrophy Type 2.” *PLoS ONE* 5 (2): e9301. doi:10.1371/journal.pone.0009301.

Sarkar, Partha S., Binoy Appukuttan, Jennifer Han, Yoshihiro Ito, Cuiwei Ai, Wenli Tsai, Yang Chai, J. Timothy Stout, and Sita Reddy. 2000. “Heterozygous Loss of Six5 in Mice Is Sufficient to Cause Ocular Cataracts.” *Nature Genetics* 25 (1): 110–14. doi:10.1038/75500.

Scheiner, Samuel M. 1993. “Genetics and Evolution of Phenotypic Plasticity.” *Annual Review of Ecology and Systematics* 24 (1): 35–68. doi:10.1146/annurev.es.24.110193.000343.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.” *Science (New York, N.Y.)* 270 (5235): 467–70.

Schena, M., D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. 1996. “Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1000 Genes.” *Proceedings of the National Academy of Sciences* 93 (20): 10614–19.

Scheres, Sjors H. W. 2012. “RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination.” *Journal of Structural Biology* 180 (3): 519–30. doi:10.1016/j.jsb.2012.09.006.

Schlichting, Carl D., and Donald A. Levin. 1984. “Phenotypic Plasticity of Annual Phlox: Tests of Some Hypotheses.” *American Journal of Botany* 71 (2): 252–60. doi:10.2307/2443753.

Schlichting, C. D., and M. Pigliucci. 1993. “Control of Phenotypic Plasticity via Regulatory Genes.” *The American Naturalist* 142 (2): 366–70. doi:10.1086/285543.

Schoen, E. D. 1996. “Statistical Designs in Combination Toxicology: A Matter of Choice.” *Food and Chemical Toxicology*, Combination Toxicology, 34 (11–12): 1059–65. doi:10.1016/S0278-6915(97)00075-6.

Schoser, Benedikt, and Lubov Timchenko. 2010. “Myotonic Dystrophies 1 and 2: Complex Diseases with Complex Mechanisms.” *Current Genomics* 11 (2): 77–90. doi:10.2174/138920210790886844.

Schüller, Hans-Joachim. 2003. “Transcriptional Control of Nonfermentative Metabolism in the Yeast Saccharomyces Cerevisiae.” *Current Genetics* 43 (3): 139–60. doi:10.1007/s00294-003-0381-8.

Senko, Michael W., Philip M. Remes, Jesse D. Canterbury, Raman Mathur, Qingyu Song, Shannon M. Eliuk, Chris Mullen, et al. 2013. “Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates.” *Analytical Chemistry* 85 (24): 11710–14. doi:10.1021/ac403115c.

Seznec, Hervé, Onnik Agbulut, Nicolas Sergeant, Cédric Savouret, Antoine Ghestem, Nacira Tabti, Jean-Claude Willer, et al. 2001. “Mice Transgenic for the Human Myotonic Dystrophy Region with Expanded CTG Repeats Display Muscular and Brain Abnormalities.” *Human Molecular Genetics* 10 (23): 2717–26. doi:10.1093/hmg/10.23.2717.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504. doi:10.1101/gr.1239303.

Shatkin, A. J. 1976. “Capping of Eucaryotic mRNAs.” *Cell* 9 (4): 645–53. doi:10.1016/0092-8674(76)90128-8.

Shaw, D. J., A. L. Meredith, M. Sarfarazi, H. G. Harley, S. M. Huson, J. D. Brook, L. Bufton, M. Litt, T. Mohandas, and P. S. Harper. 1986. “Regional Localisations and Linkage Relationships of Seven RFLPs and Myotonic Dystrophy on Chromosome 19.” *Human Genetics* 74 (3): 262–66. doi:10.1007/BF00282545.

Shelbourne, Peggy, Robert Winqvist, Erich Kunert, June Davies, Jaakko Leisti, Hannelore Thiele, Hans Bachmann, Jessica Buxton, Bob Williamson, and Keith Johnson. 1992. “Unstable DNA May Be Responsible for the Incomplete Penetrance of the Myotonic Dystrophy Phenotype.” *Human Molecular Genetics* 1 (7): 467–73. doi:10.1093/hmg/1.7.467.

Sherton, Corinne C., and Ira G. Wool. 1974. “A Comparison of the Proteins of Rat Skeletal Muscle and Liver Ribosomes by Two-Dimensional Polyacrylamide Gel Electrophoresis OBSERVATIONS ON THE PARTITION OF PROTEINS BETWEEN RIBOSOMAL SUBUNITS AND A DESCRIPTION OF TWO ACIDIC PROTEINS IN THE LARGE SUBUNIT.” *Journal of Biological Chemistry* 249 (7): 2258–67.

Sicot, Géraldine, and Mário Gomes-Pereira. 2013. “RNA Toxicity in Human Disease and Animal Models: From the Uncovering of a New Mechanism to the Development of Promising Therapies.” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, Animal models of disease, 1832 (9): 1390–1409. doi:10.1016/j.bbadis.2013.03.002.

Silver, Debra L., Dawn E. Watkins-Chow, Karisa C. Schreck, Tarran J. Pierfelice, Denise M. Larson, Anthony J. Burnetti, Hung-Jiun Liaw, et al. 2010. “The Exon Junction Complex Component Magoh Controls Brain Size by Regulating Neural Stem Cell Division.” *Nature Neuroscience* 13 (5): 551–58. doi:10.1038/nn.2527.

Sk, Jang, Pestova Tv, Hellen Cu, Witherell Gw, and Wimmer E. 1989. “Cap-Independent Translation of Picornavirus RNAs: Structure and Function of the Internal Ribosomal Entry Site.” *Enzyme* 44 (1-4): 292–309.

Slavov, Nikolai, and David Botstein. 2011. “Coupling among Growth Rate Response, Metabolic Cycle, and Cell Division Cycle in Yeast.” *Molecular Biology of the Cell* 22 (12): 1997–2009. doi:10.1091/mbc.E11-02-0132.

Smeets, Hubert, Linda Bachinski, Marga Coerwinkel, Jan Schepens, Jan Hoeijmakers, Marcel van Duin, Karl-Heinz Grzeschik, et al. 1990. “A Long-Range Restriction Map of the Human Chromosome 19q13 Region: Close Physical Linkage between CKMM and the ERCC1 and ERCC2 Genes.” *American Journal of Human Genetics* 46 (3): 492–501.

Smith, Richard D. 2002. “Trends in Mass Spectrometry Instrumentation for Proteomics.” *Trends in Biotechnology* 20 (12): s3–7. doi:10.1016/S1471-1931(02)00197-0.

Sonenberg, Nahum, and Alan G. Hinnebusch. 2009. “Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets.” *Cell* 136 (4): 731–45. doi:10.1016/j.cell.2009.01.042.

Song, Chang W. 1984. “Effect of Local Hyperthermia on Blood Flow and Microenvironment: A Review.” *Cancer Research* 44 (10 Supplement): 4721s – 4730s.

Soufi, Boumediene, Christian D. Kelstrup, Gabriele Stoehr, Florian Fröhlich, Tobias C. Walther, and Jesper V. Olsen. 2009. “Global Analysis of the Yeast Osmotic Stress Response by Quantitative Proteomics.” *Molecular BioSystems* 5 (11): 1337–46. doi:10.1039/B902256B.

Spence, Jean, Rayappa Reddy Gali, Gunnar Dittmar, Fred Sherman, Michael Karin, and Daniel Finley. 2000. “Cell Cycle–Regulated Modification of the Ribosome by a Variant Multiubiquitin Chain.” *Cell* 102 (1): 67–76. doi:10.1016/S0092-8674(00)00011-8.

Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. “BioGRID: A General Repository for Interaction Datasets.” *Nucleic Acids Research* 34 (suppl 1): D535–39. doi:10.1093/nar/gkj109.

Stastna, Miroslava, and Jennifer E. Van Eyk. 2012. “Analysis of Protein Isoforms: Can We Do It Better?” *PROTEOMICS* 12 (19-20): 2937–48. doi:10.1002/pmic.201200161.

St Johnston, Daniel. 2002. “THE ART AND DESIGN OF GENETIC SCREENS: DROSOPHILA MELANOGASTER.” *Nature Reviews Genetics* 3 (3): 176–88. doi:10.1038/nrg751.

Stuart, Joshua M., Eran Segal, Daphne Koller, and Stuart K. Kim. 2003. “A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.” *Science* 302 (5643): 249–55. doi:10.1126/science.1087447.

Suloway, Christian, James Pulokas, Denis Fellmann, Anchi Cheng, Francisco Guerra, Joel Quispe, Scott Stagg, Clinton S. Potter, and Bridget Carragher. 2005. “Automated Molecular Microscopy: The New Leginon System.” *Journal of Structural Biology* 151 (1): 41–60. doi:10.1016/j.jsb.2005.03.010.

Suzuki, Miho M., and Adrian Bird. 2008. “DNA Methylation Landscapes: Provocative Insights from Epigenomics.” *Nature Reviews Genetics* 9 (6): 465–76. doi:10.1038/nrg2341.

Tai, Siew Leng, Viktor M. Boer, Pascale Daran-Lapujade, Michael C. Walsh, Johannes H. de Winde, Jean-Marc Daran, and Jack T. Pronk. 2005. “Two-Dimensional Transcriptome Analysis in Chemostat Cultures COMBINATORIAL EFFECTS OF OXYGEN AVAILABILITY AND MACRONUTRIENT LIMITATION IN SACCHAROMYCES CEREVISIAE.” *Journal of Biological Chemistry* 280 (1): 437–47. doi:10.1074/jbc.M410573200.

Tanaka, Koichi, Hiroaki Waki, Yutaka Ido, Satoshi Akita, Yoshikazu Yoshida, Tamio Yoshida, and T. Matsuo. 1988. “Protein and Polymer Analyses up to M/z 100 000 by Laser Ionization Time-of-Flight Mass Spectrometry.” *Rapid Communications in Mass Spectrometry* 2 (8): 151–53. doi:10.1002/rcm.1290020802.

Temmerman, Nele De, Karen Sermon, Sara Seneca, Martine De Rycke, Pierre Hilven, Willy Lissens, André Van Steirteghem, and Inge Liebaers. 2004. “Intergenerational Instability of the Expanded CTG Repeat in the DMPK Gene: Studies in Human Gametes and Preimplantation Embryos.” *The American Journal of Human Genetics* 75 (2): 325–29. doi:10.1086/422762.

Terman, Jonathan R, and Anna Kashina. 2013. “Post-Translational Modification and Regulation of Actin.” *Current Opinion in Cell Biology*, Cell architecture, 25 (1): 30–38. doi:10.1016/j.ceb.2012.10.009.

The UniProt Consortium. 2015. “UniProt: A Hub for Protein Information.” *Nucleic Acids Research* 43 (D1): D204–12. doi:10.1093/nar/gku989.

Thomas, George, Jorge Martin-Perez, Michel Siegmann, and Angela M. Otto. 1982. “The Effect of Serum, EGF, PGF2α and Insulin on S6 Phosphorylation and the Initiation of Protein and DNA Synthesis.” *Cell* 30 (1): 235–42. doi:10.1016/0092-8674(82)90029-0.

Thompson, Andrew, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. 2003. “Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS.” *Analytical Chemistry* 75 (8): 1895–1904. doi:10.1021/ac0262560.

Thornton, Charles A. 2014. “Myotonic Dystrophy.” *Neurologic Clinics* 32 (3): 705–19. doi:10.1016/j.ncl.2014.04.011.
Thornton, Charles A., Robert C. Griggs, and Richard T. Moxley. 1994. “Myotonic Dystrophy with No Trinucleotide Repeat Expansion.” *Annals of Neurology* 35 (3): 269–72. doi:10.1002/ana.410350305.
Thornton, Charles A., Keith Johnson, and Richard T. Moxley. 1994. “Myotonic Dystrophy Patients Have Larger CTG Expansions in Skeletal Muscle than in Leukocytes.” *Annals of Neurology* 35 (1): 104–7. doi:10.1002/ana.410350116.
Timchenko, L. T., N. A. Timchenko, C. T. Caskey, and R. Roberts. 1996. “Novel Proteins with Binding Specificity for DNA CTG Repeats And RNA Cug Repeats: Implications for Myotonic Dystrophy.” *Human Molecular Genetics* 5 (1): 115–21. doi:10.1093/hmg/5.1.115.
Timchenko, Nikolai A., Alana L. Lu, Alana L. Lu, and Lubov T. Timchenko. 1999. “CUG Repeat Binding Protein (CUGBP1) Interacts with the 5′ Region of C/EBPβ mRNA and Regulates Translation of C/EBPβ Isoforms.” *Nucleic Acids Research* 27 (22): 4517–25. doi:10.1093/nar/27.22.4517.
Tonsor, Stephen J., Tarek W. Elnaccash, and Samuel M. Scheiner. 2013. “Developmental Instability Is Genetically Correlated with Phenotypic Plasticity, Constraining Heritability, and Fitness.” *Evolution; International Journal of Organic Evolution* 67 (10): 2923–35. doi:10.1111/evo.12175.
Trédan, Olivier, Carlos M. Galmarini, Krupa Patel, and Ian F. Tannock. 2007. “Drug Resistance and the Solid Tumor Microenvironment.” *Journal of the National Cancer Institute* 99 (19): 1441–54. doi:10.1093/jnci/djm135.
Triantafilou, Kathy, Martha Triantafilou, and Russell L. Dedrick. 2001. “A CD14-Independent LPS Receptor Cluster.” *Nature Immunology* 2 (4): 338–45. doi:10.1038/86342.
Turner, Chris, and David Hilton-Jones. 2014. “Myotonic Dystrophy: Diagnosis, Management and New Therapies.” *Current Opinion in Neurology* 27 (5): 599–606. doi:10.1097/WCO.0000000000000128.
Turowski, Tomasz W., and David Tollervey. 2014. “Cotranscriptional Events in Eukaryotic Ribosome Synthesis.” *Wiley Interdisciplinary Reviews: RNA*, August, n/a – n/a. doi:10.1002/wrna.1263.
Udd, Bjarne, and Ralf Krahe. 2012. “The Myotonic Dystrophies: Molecular, Clinical, and Therapeutic Challenges.” *The Lancet Neurology* 11 (10): 891–905. doi:10.1016/S1474-4422(12)70204-1.
Udd, Bjarne, Ralf Krahe, Carina Wallgren-Pettersson, Björn Falck, and Hannu Kalimo. 1997. “Proximal Myotonic Dystrophy—a Family with Autosomal Dominant Muscular Dystrophy, Cataracts, Hearing Loss and Hypogonadism: Heterogeneity of Proximal Myotonic Syndromes?” *Neuromuscular Disorders* 7 (4): 217–28. doi:10.1016/S0960-8966(97)00041-2.
Ulane, Christina M., Sarah Teed, and Jacinda Sampson. 2014. “Recent Advances in Myotonic Dystrophy Type 2.” *Current Neurology and Neuroscience Reports* 14 (2): 1–7. doi:10.1007/s11910-013-0429-1.
Usuki, F., S. Ishiura, N. Saitoh, N. Sasagawa, H. Sorimachi, H. Kuzume, K. Maruyama, T. Terao, and K. Suzuki. 1997. “Expanded CTG Repeats in Myotonin Protein Kinase Suppresses Myogenic Differentiation.” *Neuroreport* 8 (17): 3749–53.
Vaga, Stefania, Marti Bernardo-Faura, Thomas Cokelaer, Alessio Maiolica, Christopher A. Barnes, Ludovic C. Gillet, Björn Hegemann, et al. 2014. “Phosphoproteomic Analyses Reveal Novel Cross-modulation Mechanisms between Two Signaling Pathways in Yeast.” *Molecular Systems Biology* 10 (12): 767. doi:10.15252/msb.20145112.

van Riggelen, Jan, Alper Yetil, and Dean W. Felsher. 2010. “MYC as a Regulator of Ribosome Biogenesis and Protein Synthesis.” *Nature Reviews Cancer* 10 (4): 301–9. doi:10.1038/nrc2819.
Vaupel, Peter, Friedrich Kallinowski, and Paul Okunieff. 1989. “Blood Flow, Oxygen and Nutrient Supply, and Metabolic Microenvironment of Human Tumors: A Review.” *Cancer Research* 49 (23): 6449–65.
Veeramah, Krishna R., Laurel Johnstone, Tatiana M. Karafet, Daniel Wolf, Ryan Sprissler, John Salogiannis, Asa Barth-Maron, et al. 2013. “Exome Sequencing Reveals New Causal Mutations in Children with Epileptic Encephalopathies.” *Epilepsia* 54 (7): 1270–81. doi:10.1111/epi.12201.
Venema, Jaap, and David Tollervey. 1999. “Ribosome Synthesis in Saccharomyces Cerevisiae.” *Annual Review of Genetics* 33 (1): 261–311. doi:10.1146/annurev.genet.33.1.261.
Via, Sara, and Russell Lande. 1985. “Genotype-Environment Interaction and the Evolution of Phenotypic Plasticity.” *Evolution* 39 (3): 505–22. doi:10.2307/2408649.
Vincent P Mauro, and Gerald M Edelman. 2007. “The Ribosome Filter Redux.” *Cell Cycle* 6 (18): 2246–51.
Visser, J. Arjan G. M. de, Tim F. Cooper, and Santiago F. Elena. 2011. “The Causes of Epistasis.” *Proceedings of the Royal Society B: Biological Sciences* 278 (1725): 3617–24. doi:10.1098/rspb.2011.1537.
Voermans, N. C., A. E. Laan, A. Oosterhof, T. H. van Kuppevelt, G. Drost, M. Lammens, E. J. Kamsteeg, et al. 2012. “Brody Syndrome: A Clinically Heterogeneous Entity Distinct from Brody Disease: A Review of Literature and a Cross-Sectional Clinical Study in 17 Patients.” *Neuromuscular Disorders* 22 (11): 944–54. doi:10.1016/j.nmd.2012.03.012.
Volarević, Siniša, and George Thomas. 2000. “Role of S6 Phosphorylation and S6 Kinase in Cell Growth.” In , edited by BT - Progress in Nucleic Acid Research and Molecular Biology, 65:101–27. Academic Press. http://www.sciencedirect.com/science/article/pii/S0079660300650031.
Vucic, Domagoj, Vishva M. Dixit, and Ingrid E. Wertz. 2011. “Ubiquitylation in Apoptosis: A Post-Translational Modification at the Edge of Life and Death.” *Nature Reviews Molecular Cell Biology* 12 (7): 439–52. doi:10.1038/nrm3143.
Walther, Tobias C., and Matthias Mann. 2010. “Mass Spectrometry–based Proteomics in Cell Biology.” *The Journal of Cell Biology* 190 (4): 491–500. doi:10.1083/jcb.201004052.
Wang, Eric T., Neal A. L. Cody, Sonali Jog, Michela Biancolella, Thomas T. Wang, Daniel J. Treacy, Shujun Luo, et al. 2012. “Transcriptome-Wide Regulation of Pre-mRNA Splicing and mRNA Localization by Muscleblind Proteins.” *Cell* 150 (4): 710–24. doi:10.1016/j.cell.2012.06.041.
Warner, Jonathan R. 1999. “The Economics of Ribosome Biosynthesis in Yeast.” *Trends in Biochemical Sciences* 24 (11): 437–40. doi:10.1016/S0968-0004(99)01460-7.
Warner, Jonathan R. 2015. “Twenty Years of Ribosome Assembly and Ribosomopathies.” *RNA* 21 (4): 758–59. doi:10.1261/rna.050435.115.
Warner, Jonathan R., and Kerri B. McIntosh. 2009. “How Common Are Extraribosomal Functions of Ribosomal Proteins?” *Molecular Cell* 34 (1): 3–11. doi:10.1016/j.molcel.2009.03.006.
Warner, J. R. 1989. “Synthesis of Ribosomes in Saccharomyces Cerevisiae.” *Microbiological Reviews* 53 (2): 256–71.
Wells, Lance, Stephen A Whelan, and Gerald W Hart. 2003. “O-GlcNAc: A Regulatory Post-Translational Modification.” *Biochemical and Biophysical Research Communications* 302 (3): 435–41. doi:10.1016/S0006-291X(03)00175-X.
Whiteside, T. L. 2008. “The Tumor Microenvironment and Its Role in Promoting Tumor Growth.” *Oncogene* 27 (45): 5904–12. doi:10.1038/onc.2008.271.

Wilhelm, Brian T., Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J. Penkett, Jane Rogers, and Jürg Bähler. 2008. “Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution.” *Nature* 453 (7199): 1239–43. doi:10.1038/nature07002.

Wilhelm, Mathias, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, et al. 2014. “Mass-Spectrometry-Based Draft of the Human Proteome.” *Nature* 509 (7502): 582–87. doi:10.1038/nature13319.

Wilson, M., and S. E. Lindow. 1993. “Effect of Phenotypic Plasticity on Epiphytic Survival and Colonization by Pseudomonas Syringae.” *Applied and Environmental Microbiology* 59 (2): 410–16.

Winter, Georg E., Uwe Rix, Scott M. Carlson, Karoline V. Gleixner, Florian Grebien, Manuela Gridling, André C. Müller, et al. 2012. “Systems-Pharmacology Dissection of a Drug Synergy in Imatinib-Resistant CML.” *Nature Chemical Biology* 8 (11): 905–12. doi:10.1038/nchembio.1085.

Winzeler, Elizabeth A., Daniel D. Shoemaker, Anna Astromoff, Hong Liang, Keith Anderson, Bruno Andre, Rhonda Bangham, et al. 1999. “Functional Characterization of the S. Cerevisiae Genome by Gene Deletion and Parallel Analysis.” *Science* 285 (5429): 901–6. doi:10.1126/science.285.5429.901.

Wold, F. 1981. “In Vivo Chemical Modification of Proteins (Post-Translational Modification).” *Annual Review of Biochemistry* 50 (1): 783–814. doi:10.1146/annurev.bi.50.070181.004031.

Woolford, John. 2015. “Assembly of Ribosomes in Eukaryotes.” *RNA* 21 (4): 766–68. doi:10.1261/rna.050633.115.

Woolford, John L., and Susan J. Baserga. 2013. “Ribosome Biogenesis in the Yeast Saccharomyces Cerevisiae.” *Genetics* 195 (3): 643–81. doi:10.1534/genetics.113.153197.

Wulfkuhle, Julia D., Lance A. Liotta, and Emanuel F. Petricoin. 2003. “Proteomic Applications for the Early Detection of Cancer.” *Nature Reviews Cancer* 3 (4): 267–75. doi:10.1038/nrc1043.

Wu, Tsung-Jung, Amirhossein Shamsaddini, Yang Pan, Krista Smith, Daniel J. Crichton, Vahan Simonyan, and Raja Mazumder. 2014. “A Framework for Organizing Cancer-Related Variations from Existing Databases, Publications and NGS Data Using a High-Performance Integrated Virtual Environment (HIVE).” *Database* 2014 (January): bau022. doi:10.1093/database/bau022.

Xiong, X., Y. Zhao, H. He, and Y. Sun. 2011. “Ribosomal Protein S27-like and S27 Interplay with p53-MDM2 Axis as a Target, a Substrate and a Regulator.” *Oncogene* 30 (15): 1798–1811. doi:10.1038/onc.2010.569.

Xirodimas, Dimitris P, Anders Sundqvist, Akihiro Nakamura, Linnan Shen, Catherine Botting, and Ronald T Hay. 2008. “Ribosomal Proteins Are Targets for the NEDD8 Pathway.” *EMBO Reports* 9 (3): 280–86. doi:10.1038/embor.2008.10.

Xu, Danmei, Naoko Suenaga, Mariola J. Edelmann, Rafael Fridman, Ruth J. Muschel, and Benedikt M. Kessler. 2008. “Novel MMP-9 Substrates in Cancer Cells Revealed by a Label-Free Quantitative Proteomics Approach.” *Molecular & Cellular Proteomics* 7 (11): 2215–28. doi:10.1074/mcp.M800095-MCP200.

Xue, Shifeng, and Maria Barna. 2012. “Specialized Ribosomes: A New Frontier in Gene Regulation and Organismal Biology.” *Nature Reviews Molecular Cell Biology* 13 (6): 355–69. doi:10.1038/nrm3359.

Xue, Shifeng, Siqi Tian, Kotaro Fujii, Wipapat Kladwang, Rhiju Das, and Maria Barna. 2015. “RNA Regulons in Hox 5′ UTRs Confer Ribosome Specificity to Gene Regulation.” *Nature* 517 (7532): 33–38. doi:10.1038/nature14010.

Yahata, Tetsuro, Mark P. de Caestecker, Robert J. Lechleider, Stephanie Andriole, Anita B. Roberts, Kurt J. Isselbacher, and Toshi Shioda. 2000. “The MSG1 Non-DNA-Binding Transactivator Binds to the p300/CBP Coactivators, Enhancing Their Functional Link to the Smad Transcription Factors.” *Journal of Biological Chemistry* 275 (12): 8825–34. doi:10.1074/jbc.275.12.8825.

Yang, Zongde, Xin Chen, Qiulin Zhang, Bin Cai, Kai Chen, Ziqiang Chen, Yushu Bai, Zhicai Shi, and Ming Li. 2015. “Dysregulated COL3A1 and RPL8, RPS16, RPS23 in Disc Degeneration Revealed by Bioinformatics Methods.” *Spine*, April. doi:10.1097/BRS.0000000000000939.

Yan, Shun-Ping, Qun-Ye Zhang, Zhang-Cheng Tang, Wei-Ai Su, and Wei-Ning Sun. 2006. “Comparative Proteomic Analysis Provides New Insights into Chilling Stress Responses in Rice.” *Molecular & Cellular Proteomics* 5 (3): 484–96. doi:10.1074/mcp.M500251-MCP200.

Yates, John R., Cristian I. Ruse, and Aleksey Nakorchevsky. 2009. “Proteomics by Mass Spectrometry: Approaches, Advances, and Applications.” *Annual Review of Biomedical Engineering* 11 (1): 49–79. doi:10.1146/annurev-bioeng-061008-124934.

Yelick, Pamela C., and Paul A. Trainor. 2015. “Ribosomopathies: Global Process, Tissue Specific Defects.” *Rare Diseases* 3 (1): e1025185. doi:10.1080/21675511.2015.1025185.

Young, David J., Nicholas R. Guydosh, Fan Zhang, Alan G. Hinnebusch, and Rachel Green. 2015. “Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3′UTRs In Vivo.” *Cell* 162 (4): 872–84. doi:10.1016/j.cell.2015.07.041.

Young, Sara K., Jeffrey A. Willy, Cheng Wu, Matthew S. Sachs, and Ronald C. Wek. 2015. “Ribosome Reinitiation Directs Gene-Specific Translation and Regulates the Integrated Stress Response.” *Journal of Biological Chemistry*, October, jbc.M115.693184. doi:10.1074/jbc.M115.693184.

Zanni, Ginevra, Vera M. Kalscheuer, Andreas Friedrich, Sabina Barresi, Paolo Alfieri, Matteo Di Capua, Stefan A. Haas, et al. 2015. “A Novel Mutation in RPL10 (Ribosomal Protein L10) Causes X-Linked Intellectual Disability, Cerebellar Hypoplasia and Spondylo-Epiphyseal Dysplasia.” *Human Mutation*, August, n/a – n/a. doi:10.1002/humu.22860.

Zemp, Ivo, and Ulrike Kutay. 2007. “Nuclear Export and Cytoplasmic Maturation of Ribosomal Subunits.” *FEBS Letters*, Vienna Special Issue: Molecular Machines, 581 (15): 2783–93. doi:10.1016/j.febslet.2007.05.013.

Zhang, Bin, and Steve Horvath. 2005. “A General Framework for Weighted Gene Co-Expression Network Analysis.” *Statistical Applications in Genetics and Molecular Biology* 4 (1). doi:10.2202/1544-6115.1128.

Zhang, Yaoyang, Bryan R. Fonslow, Bing Shan, Moon-Chang Baek, and John R. Yates. 2013. “Protein Analysis by Shotgun/Bottom-up Proteomics.” *Chemical Reviews* 113 (4): 2343–94. doi:10.1021/cr3003533.

Zhao, Yong, Joshua F. Ransom, Ankang Li, Vasanth Vedantham, Morgan von Drehle, Alecia N. Muth, Takatoshi Tsuchihashi, Michael T. McManus, Robert J. Schwartz, and Deepak Srivastava. 2007. “Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking miRNA-1-2.” *Cell* 129 (2): 303–17. doi:10.1016/j.cell.2007.03.030.