GENE EXPRESSION AND COPY NUMBER VARIATION IN COMMON COMPLEX DISEASES

By

Britney L. Grayson

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

Microbiology and Immunology

December, 2010

Nashville, Tennessee

Approved:

Thomas M. Aune

Luc Van Kaer

James W. Thomas

Nancy J. Brown

Marylyn D. Ritchie

To my mom, Ruthie, your battle with lupus inspired every experiment on these pages.

To my dad, Richard, a short pencil IS better than a long memory.

To my sisters, Brandy and Brienne, I am but one-third of a whole.

ACKNOWLEDGEMENTS

I am thankful to my thesis committee, especially my mentor Tom Aune for his many hours of "interaction" and countless "pearls of wisdom." I am thankful to Luc Van Kaer for being an organized and timely chair and to all members, Drs. Aune, Van Kaer, Thomas, Ritchie and Brown, for insightful comments, critiques and advice that propelled both me and these projects forward. I am thankful for my lab members past and present- Kelly, Patrick, Zach, Erica, Sarah, Nyk, Chase, our honorary member, David and especially Mary Ellen, who performed many of the experiments you see on these pages, John, who helps to maintain our patient samples biobank and Mel, who had a solution to all of my lab problems.

I am thankful to all of the physicians and clinics who gave of their time to recruit patients for these studies. This includes Drs. James W. Thomas, Howard Fuchs, Nancy J. Brown, Joe Huston, Bill Russell and Steve Davis, Margo Black and the Eskind Diabetes Clinic, and the nurses at the Clinical Research Center at Vanderbilt. I am also thankful to Dr. George Eisenbarth, Joy Jeffrey, Pam Fain and Priyaanka Nanduri at the Barbara Davis Center for Childhood Diabetes at the University of Colorado for collaborating with me on the work presented in Chapter IV.

I would also like to acknowledge my many collaborators outside of the clinics: The Vanderbilt Functional Genomics Shared Resource, specifically Vicky Amann, Braden Boone and Phil Dexheimer for technical support with arrays both in experimental procedures and statistical analysis, and Latha Raju for her technical expertise in RT-PCR. My work would not be as "significant" without the help of our wonderful statistician,

Lily Wang, and the Vanderbilt Department of Biostatistics. Every aspect of my work was made stronger by her tireless work.

Thank you to my most important collaborators- the hundreds of patients who made this work possible.

I am thankful to the MSTP, our director Terry Dermody and the rest of the team- Larry, Jim, Michelle, Susan, Jena, Gisel and Lindsay for unending support since my interview weekend nearly 7 years ago. Thanks to the Department of Microbiology and Immunology, chair Jacek Hawiger, administrator extraordinaire Jean Tidwell, director of graduate studies Chris Aiken, and the numerous others who have supported my graduate training for the last 4 years.

On a personal note, I am so thankful to my family and friends. To all my MSTP colleagues, especially Aubrey and Katy, for understanding the highs and lows of graduate school life. Thanks to my Nashville friends, who are like family, for supporting every day of my crazy life- whether I could reciprocate or not. Finally, thank you to my family. In addition to those for whom this work is dedicated, I would like to thank Alecia, Angel, Nicole, Grandpa, Pop, Gram and the rest of the Noble gang for their unconditional love and support of my work.

Financially, this work was supported by NIH grants R42 AI053984, T32 GM07347 (Medical Scientist Training Program), T32 DK07563 (Molecular Endocrinology Training Grant), and TL1 RR024978 (Clinical and Translational Research Award).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAD- coronary artery disease

CGH- comparative genomic hybridization

CN- copy number

CNP- copy number polymorphism

CNV- copy number variation

CNVR- copy number variant region

CTRL- control

FISH- fluorescence in-situ hybridization

GTC- genotyping console

HEEBO- human exonic evidence-based oligonucleotide

HLA- human leukocyte antigen

MetS- metabolic syndrome

MHC- major histocompatibility complex

MS- multiple sclerosis

qPCR- quantitative polymerase chain reaction

RA- rheumatoid arthritis

RT-PCR- reverse transcriptase-polymerase chain reaction

SLE- systemic lupus erythematosus

SNP- single nucleotide polymorphism

T1D- type 1 diabetes

T2D- type 2 diabetes

CHAPTER I

INTRODUCTION

Common, complex diseases are not caused by a single gene mutation but rather have genetic and environmental components. As a subset, autoimmune diseases also possess robust gene expression signatures that have both genetic and environmental contributions. We explored gene expression signatures in type 2 diabetes, coronary artery disease, and their precursor state metabolic syndrome, and identified overlapping gene expression signatures exhibiting greater resemblance to each other than to an autoimmune disease. These signatures are consistent with activation of the innate immune response. Genetic variations contribute to familiality and we sought to determine if large-scale genomic variants, copy number variants (CNVs), are associated with common, complex diseases. To do so, we first developed methods to identify CNVs from SNP-based arrays. We analyzed genomic variation in type 1 diabetes and identified CNVs that were differentially present in patients with or at high risk for type 1 diabetes versus control. Thus, we conclude that both gene expression profiles and genomic variants are easily detected clinical markers that may be useful predictors of disease liability and serve to identify new classes of therapeutic targets.

Common Complex Disease

Complex diseases can include all those not known to be caused by a single gene mutation. Rather, they are thought to be caused by a combination of multiple genetic and environmental factors. Examples include chronic obstructive pulmonary disease, Parkinson disease, hypertension and diabetes. Autoimmune disease is one class of complex diseases with a strong genetic component. This group of diseases affects 3-5% of the human population[1] and can be further divided into diseases that are systemic, like systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) and those that target a single organ, like type 1 diabetes (T1D) and multiple sclerosis (MS).

Another class of common complex diseases includes such conditions as type 2 diabetes (T2D), coronary artery disease (CAD) and a well-defined precursor state to both T2D and CAD, the metabolic syndrome (MetS). These conditions will be called "metabolic disorders" because obesity is a trait strongly associated with all three. Obesity is also increasingly common in the United States. The U.S. prevalence of MetS may be as high as 39%, indicating that over 1/3 of the nation is at high risk for developing T2D or CAD, and possibly both[2,3].

Autoimmune Disease

SLE is a systemic autoimmune disease where antibodies to the components of cell nuclei can be detected[4]. These autoantibodies are able to bind in any cell, tissue and organ in the body, often causing widespread disease. RA is a disease of joint inflammation triggered by lymphocytic and monocytic infiltration to the synovium,

synovial hyperplasia, antibodies binding to the synovium, joint destruction and systemic inflammation[5]. T1D results from immune-mediated selective destruction of insulin producing beta cells of the pancreatic islets resulting in insulin deficiency and hyperglycemia[6,7]. Symptoms of polydipsia, polyuria, polyphagia and weight loss manifest when significant numbers of beta cells are destroyed. MS is a neurological disease characterized by demyelination of the myelin sheaths of neurons resulting in symptoms such as vision and sensorimotor disturbances[8].

A central tenet of the immune system is the ability of lymphocyte receptors to distinguish self from non-self; this concept is called tolerance. Lymphocyte specificity is determined by random recombination events in order to create a large repertoire of antigen receptors. This is achieved by recombination of one of each of three essential receptor segments: variable (V), diversity (D) and joining (J). By the nature of the random generation of these receptors, between 20 and 50% of them will recognize and bind self proteins and must be edited or removed centrally, or controlled peripherally, to prevent autoimmune disease[9]. The processes by which tolerance is established differ slightly for B cells and T cells[10].

B cells that react with self antigen can either be deleted or undergo receptor editing of the light chain during their development in the bone marrow; this is called central tolerance. Receptor editing changes the antigen specificity of the cell and ideally prevents it from binding to self. Receptor editing serves as the last chance for autoreactive B cells to escape deletion. If the receptor retains self-specificity, the cell is signaled to undergo apoptosis in a process referred to as negative selection. Sometimes self-reactive B cells escape deletion and are accidentally released to the periphery. When

these self-reactive cells encounter the antigen they recognize in the absence of co-stimulatory molecules, which are required for complete activation, they are stimulated in such a way to become permanently "anergic," or not responsive to antigen. Additionally, other B cells bind only weakly to self antigen and so escape detection and deletion. These B cells are usually nonfunctional (since their antigen binding is weak), but can become activated if the concentration of the antigen they recognize is unusually high.

T cells undergo both positive and negative selection during development in the thymus. Like B cells, T cells that react to self-antigen are deleted by negative selection. T cells must bind self-MHC in order to recognize antigen. Positive selection of T cells occurs when T cells bind self-MHC in the thymus and receive a signal to live and proceed to the periphery. Positive selection is important to ensure T cells that fail to recognize self-MHC are not released to the periphery, as they would essentially be nonfunctional. Clonal selection ensures that the mature T cell population can react with foreign antigen, but does not react with self-antigen. Inevitably, there are self antigens expressed at very low levels, or not at all, in the thymus. T cells reacting to these antigens do not have the opportunity to undergo negative selection. These cells are released to the periphery where they must be controlled through clonal deletion, anergy, active suppression or ignorance.

Clonal deletion occurs when a T cell that binds repeatedly to antigen (due to a high concentration of self antigen, for example) undergoes programmed cell death. Anergy, as previously described, is a state in which the T cell recognizes self-antigen but remains inactive due to a lack of the co-stimulatory molecules required for activation of the T cell. Self reactive T cells are kept non-functional when self antigen is presented at

low levels, a process called active suppression. The cells reacting at low levels differentiate into regulatory cells, which prevent other cells from reacting to that antigen by secreting regulatory cytokines. This is the most common mechanism for controlling potentially autoreactive T cells in the periphery. Finally, like B cells, if a self-reactive T cell never encounters its antigen, it will be harmless under normal conditions. This is called clonal ignorance.

In all autoimmune diseases, there is a fundamental loss of tolerance that develops through one or more of several possible pathways. Mechanisms proposed to account for loss of tolerance to self-antigen include, but are not limited to, homeostatic expansion of autoreactive T cells, cross-reactivity of infectious antigens with self-proteins, and genetic predisposition to recognition of self-antigens.

Homeostatic expansion is the proliferation of certain subtypes of T cells in response to relative lymphopenia, a process driven both by lack of T cells in the lymphocyte compartment and cytokine growth factors[11,12,13]. Lymphopenia, due to thymectomy or a viral infection, for instance, is often found as a precursor to the development of autoimmunity[14] and as a feature of certain autoimmune diseases, like T1D[15]. It is hypothesized that in the T cell depleted state, autoreactive T cells preferentially expand and trigger the onset of autoimmune disease[14,16].

Cross reactivity of an antibody, originally generated to fight an infection, with a self-antigen could also trigger development of an autoimmune disease. For example, in MS, antibodies are found that react with the nervous system protein myelin[17]. Sequencing of several viral genomes including influenza, measles and Epstein-Barr virus show that each of these genomes contain sequences similar to those found in myelin, a concept

called “molecular mimicry.” During an infection with any of these viruses, there is a chance that an antibody will be generated that is cross reactive with myelin, potentially resulting in the onset of MS.

Finally, genetics are also known to play a role in the loss of tolerance and occurrence of autoimmune disease. One of the initial observations that suggested a genetic component was that this group of diseases tends to cluster in families and be present generation after generation. One way to study the inheritance of disease is to determine incidences in monozygotic and dizygotic twin pairs. Presumably, monozygotic twins share the exact same genome while dizygotic twins share only as much as non-twin siblings do, estimated to be no more than 50% similar. By studying pairs in which one twin is diagnosed with an autoimmune disease and counting the frequency with which the co-twin is also diagnosed with that disease, the impact of genetics can be assessed.

The incidence of RA in a monozygotic co-twin of a diagnosed subject is 15.4% whereas the incidence of disease in a dizygotic co-twin is only 3.6%[18]. Data from twin studies also make it possible to estimate the heritability of RA by accounting for factors that may compound the data, like age, gender and clinical disease characteristics. Based upon two separate analyses, the genetic component of RA ranges from 53-65%[19]. T1D shows similar genetic heritability. Monozygotic twins have anywhere from 27.3-65% concordance rate while the disease concordance rate for dizygotic twins is only 3.8%[20,21]. Interestingly, the average time to diagnosis of T1D in the second twin is also shorter (6.9 years) in monozygotic than dizygotic twins (23.6 years)[20]. Additional studies have tracked the unaffected twin in a T1D-discordant pair not only for development of diabetes, but for presence of an islet specific autoantibody in serum. While cumulative

incidence of diabetes in the unaffected twin was 65% by age 60, persistent autoantibody positivity was detected in 78% of the unaffected twins in the same time frame[21].

The major histocompatibility complex (MHC) is a highly polymorphic region of the genome, found on chromosome 6, that encodes proteins (human leukocyte antigens, HLA) responsible for the presentation of antigens to immune system cells. HLA types are inherited and certain are shown to provide protection from or susceptibility to autoimmune disease. For instance, HLA-DR and HLA-DQ confer the greatest susceptibility to developing MS; HLA-DRB1 is a risk allele for RA and the DQ allele also confers risk for T1D[17,22,23]. Additional genes have been associated with autoimmune disorders but do not carry the same impact on heritability as the MHC[24,25,26]. It has been estimated that as much as 30% of the heritability of RA is derived from HLA alleles while estimates for the heritability of T1D place the importance of the MHC at 50%[22,27].

While homeostatic expansion of T cells, cross reactivity following infection and genetics are all sound bases for the development of autoimmune diseases, none have been shown to operate independently of the other to cause disease. How these three work together to definitively trigger disease is not entirely understood.

Metabolic Disorders

Type 2 diabetes (T2D) is a metabolic disorder of peripheral insulin resistance[28]. Insulin is a pancreatic hormone involved in tightly regulating blood glucose levels. Insulin resistance occurs when the biological effect of insulin does not achieve its purpose of disposing of blood glucose in skeletal muscle and suppressing hepatic glucose production. The result, hyperglycemia, should stimulate greater insulin production and

secretion from the beta cells of the pancreas. In T2D, the physiological response to hyperglycemia is not sufficient to reduce blood glucose levels to the normal range, resulting in maintained hyperglycemia and glucose toxicity. Glucose toxicity is another mechanism that induces beta cell dysfunction. While T2D begins as a problem of insulin resistance, prolonged untreated hyperglycemia may ultimately cause partial or complete loss of beta cell function requiring pharmacologic insulin therapy. Risk factors for T2D include obesity, physical inactivity and family history. Adipocytes secrete non-esterified fatty acids that affect the ability of insulin action at effector sites like the liver and skeletal muscle. Twin studies in type 2 diabetics show the monozygotic concordance rate of T2D ranges from 35-58% while the dizygotic concordance is 17-20%[29]. Thus we can see that even in situations of presumably identical environmental exposures, genetic factors have an additional impact on the occurrence of T2D.

Coronary artery disease (CAD) results from atherosclerotic plaque development in coronary arteries[30]. A plaque begins as a so-called "fatty streak," when lipid-filled macrophages accumulate in the inner layers of the artery wall. The process that follows layers smooth muscle cells with additional lipid-filled macrophages underneath a cap of connective tissue to create a complex lesion. This fibrous, fatty plaque can ultimately block the flow of blood or rupture the vessel, resulting in angina and/or myocardial infarction. Risk factors for CAD are similar to those for T2D. Hyperlipidemia predisposes to the development of these plaques, making obesity and physical inactivity risk factors for CAD. CAD also has a genetic component as studies of risk of death due to CAD have shown the relative risk of death to be 8.1-15.0 for monozygotic twins whose

co-twin died of complications related to CAD prior to age 55, with only a 2.6-3.8 relative risk of death in dizygotic twins[31].

Metabolic syndrome (MetS) is a well-defined precursor state to both T2D and CAD[3,32,33,34,35,36,37,38]. The International Diabetes Federation (IDF) defines this pre-disease state as central obesity plus any 2 of the following 4 characteristics: hypertriglyceridemia, low levels of high-density lipoprotein (HDL) cholesterol, hypertension or raised fasting plasma glucose. These criteria reference a number of underlying biological processes. Raised fasting plasma glucose, for instance, is the first clinically detectable sign of insulin resistance or beta cell dysfunction. As previously mentioned, hyperlipidemia, including hypertriglyceridemia, is an independent risk factor for both T2D and CAD that is included in the definition of MetS.

The prevalence of MetS in the United States is as high as 39% using the IDF criteria[2]. Diagnosis of MetS confers a 1.5-2.6 relative risk of developing CAD[33,34] and a 3.5-7.5 relative risk of developing T2D[3,35,39]. Additionally, the Framingham study determined that a portion of these relative risks persist even in the absence of obesity[40]. This trio of disorders poses a significant threat to public health in the United States but one that is not irreversible. Weight loss and dietary modifications can reverse the criteria used to diagnose metabolic syndrome and decrease the patient's consequent risk for T2D and CAD.

Gene Expression Profiling

In the mid-1990s robotic printing of cDNAs onto glass slides birthed microarray technology and the ability to measure the mRNA expression levels of large numbers of genes in a biological sample in one experiment[41,42]. Soon after, these arrays were employed to define cellular and disease phenotypes, biological responses to stimuli as well as to monitor progression of disease and responses to treatment. Both organ target tissues and peripheral blood have been used as substrates for transcript isolation and microarray analysis.

Gene expression microarrays are built using oligonucleotide probes complementary to the sequences of mRNA transcripts. Early arrays measured expression based on sequences in the 3' untranslated region for targeted groups of genes numbering in the thousands. With advancing technology, it has become possible to assess the expression levels of all known genes with probes that can recognize and bind various parts of the transcript including exon-exon boundaries and alternatively spliced sequences. Two options for genome wide gene expression analysis are the Human Exonic Evidence Based Oligonucleotide, or HEEBO, array (www.microarray.org) and the Affymetrix GeneChip array (www.affymetrix.com). The HEEBO array contains nearly 45,000 cDNA probes designed by researchers at Stanford University using a transcriptome based library of exonic structure. The probes include both constitutively expressed and alternatively spliced sequences for maximum detection of all known transcripts of genes. The Affymetrix GeneChip is one of the most widely used gene

expression arrays and measures expression of nearly 29,000 genes with multiple probes per gene spanning the length of the transcript.

To analyze gene expression by microarrays, mRNA is isolated from a biologic sample and transcribed to produce fluorescently labeled cDNA. The cDNA is hybridized to the array, washed to remove excess unbound sample and scanned into a computer. The computer aligns fluorescence intensity values at each oligonucleotide probe with the corresponding gene. Two-color fluorescence can be used with a control (CTRL) sample on each array and a measurement of the ratio of intensity. Alternatively, raw intensities from each test sample can be compared to those of a separate CTRL sample or group for determination of relative expression values.

There are a number of options for analyzing microarray data, including statistical analyses offered by The Institute for Genomic Research's program TM4[43] and gene set analysis using Gene Ontology groupings[44]. TM4 is a suite of microarray analysis programs including options for normalization and viewing. The multi-experiment viewer (MeV) allows not only visualization of microarray data but also advanced statistical analyses. Statistical analysis of microarray (SAM) is a supervised comparison of two groups of microarray data that assumes gene expression is significantly different between groups[45]. Each sample is assigned to a group and SAM determines which genes are most likely to differ between those groups. This result comes with a false discovery rate estimating the number of genes likely to have been identified as differential by chance alone. The SAM output can be used as input for support tree hierarchical clustering with bootstrap re-sampling. Bootstrapping is a method of re-sampling with replacement where each group retains the same number of samples, but in each comparison certain samples

are dropped and others are duplicated. This allows the impact of an outlier sample to be minimized and support can be calculated for the similarity of samples to each other based on the significantly differentially expressed genes determined by SAM.

Additionally, gene expression data can be analyzed in the context of gene sets to increase the likelihood of discovering functional associations[46]. In this process, intensity based expression levels of all genes measured are grouped by similar function or participation in shared pathways using research deposited into public databases. Genes may belong to more than one gene set. By grouping genes, differential expression of pathways and processes can be detected on a biologically relevant level. Of note, these analyses work within the framework of known gene associations and pathways. The identification of novel gene-pathway associations would not be possible through this type of gene set analysis.

Gene Expression Profiling of Cancer

One of the first applications of microarray gene expression technology was in the field of cancer. Gene expression profiles were characterized for known tumors and consequently, transcripts isolated from potentially malignant tumors could be analyzed by microarray and the type of cancer could be determined. These tests allowed for the better typing of known classes of tumors as well as further classification of cancer types[47,48,49]. Prior knowledge of diagnosis or cancer type in a study of acute leukemia was not necessary to inform the distinctions in gene expression between types making gene expression microarray a powerful and sufficient tool for future diagnoses[50].

Certain of these expression signatures are also associated with differential outcomes. Microarray analysis of B cell lymphoma showed two distinct profiles, one most closely resembling the expression profile of germinal center B cells and the other of activated B cells[51]. Overall survival was significantly better in patients with germinal-center-like lymphoma making gene expression profiling clinically significant in the diagnosis of and care for patients with B cell lymphomas.

Gene Expression Profiling in Autoimmune Disease

Microarray analysis of transcripts isolated from peripheral whole blood can give a portrait of gene expression in lymphocyte-related diseases like autoimmune diseases[52,53]. Using an array that quantified expression of more than 4,000 genes, a gene expression signature was found to be common to RA, SLE, T1D and MS (Figure 1-1)[54]. This signature is unique from CTRL subjects and recently immunized subjects, indicating that genes differentially regulated in autoimmune disease are not those involved in a typical immune response. Rather, differentially expressed genes encode proteins involved in cell cycle regulation, differentiation, apoptosis and cell migration[55].

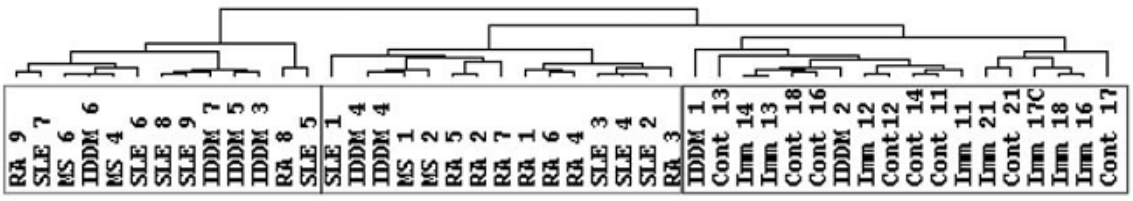

Figure 1-1. Comparison of the immune and autoimmune classes by cluster analysis. The entire data set: preimmune (Cont), postimmune (Imm), and the four autoimmune disease groups—RA, SLE, IDDM, and MS—were subjected to cluster analysis. IDDM=T1D.
Taken from: Aune, T. M., et al. "Profiles of gene expression in human autoimmune disease." Cell Biochemistry and Biophysics. 2004. 40:81-96.

To explore the reach and impact of the autoimmune gene expression signature further, transcripts isolated from peripheral blood of unaffected first degree relatives of patients with an autoimmune disease were analyzed on the same 4,000 gene microarray[56,57]. Unaffected first degree relatives showed a similar profile to their autoimmune-affected relatives, indicating that the profile is not solely representative of an active disease process (Figure 1-2). Environmental exposures or genetic inheritance may also influence expression of the genes in this signature in the absence of clinical disease[57].

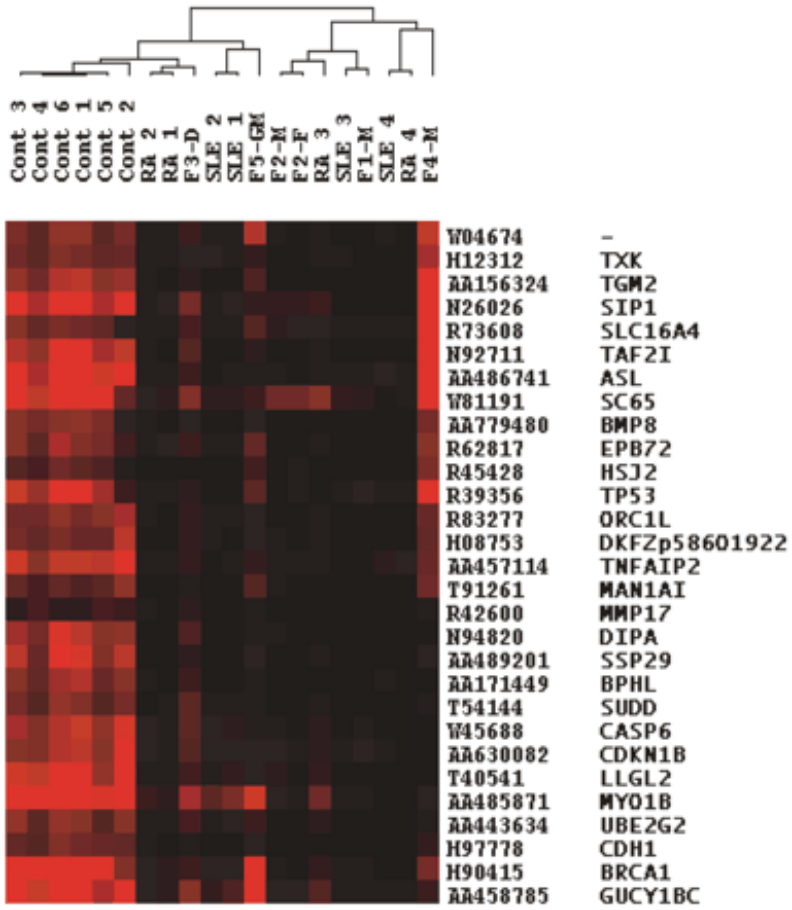

Figure 1-2. Hierarchical clustering using core autoimmune genes.
Microarray data were restricted to 29 previously identified core autoimmune genes. Profiles for control individuals (Cont), unaffected family members (F) and autoimmune individuals (RA or SLE) were subjected to hierarchical clustering. Hybridization intensities are represented as a range from black (no expression) to red (high expression). Taken from: Aune, T. M., et al. "Profiles of gene expression in human autoimmune disease." Cell Biochemistry and Biophysics. 2004. 40:81-96.

While a portion of the signature overlaps with unaffected relatives, a portion can be identified as unique to patients with RA, T1D, SLE or MS[58]. The differentially expressed genes unique to patients with an autoimmune disease fall in the broad classes of DNA damage response genes, RNA splicing, and responses to toxin.

Ability to differentiate a patient with an autoimmune disease from one without based on gene expression data derived from peripheral blood is clinically relevant when considering that certain autoimmune diseases, like SLE and MS, are notoriously difficult to diagnose and peripheral blood is an easily accessible and replenishable resource. The differentially expressed genes, together, could serve a purpose in diagnostics. In addition, the differential expression of genes in the autoimmune signature also suggests certain functional deficits might be found in lymphocytes from patients with autoimmune diseases.

The autoimmune gene expression signature showed differential expression of cell cycle response and apoptosis genes, in particular down regulation of the gene encoding the protein p53. The apoptotic response to gamma radiation is a p53-dependent cellular process. Lymphocyte viability studies showed that T cells from patients with RA had lower levels of apoptosis in response to gamma radiation than T cells from CTRL subjects[59]. This study is in agreement with research on murine models of autoimmune disease that showed defective apoptotic pathways to be involved in the pathogenesis of disease[60]. Similar gamma radiation studies in peripheral blood mononuclear cells from patients with MS showed identical deficits in apoptosis in response to radiation and further attributed the decline in p53 to a lack of stabilization of that protein by the ATM protein[61]. These studies confirmed a functional cellular deficit corresponding to altered transcript levels discovered in the peripheral blood gene expression profiles of patients with autoimmune diseases.

Additional studies have delineated disease-specific gene expression profiles for autoimmune diseases. The peripheral blood gene expression signature of SLE has been

characterized by a number of groups to include significant differential expression and dysregulation of genes involved in the interferon response[62,63,64]. The presence of this signature has also been associated with more severe disease, including kidney and nervous system involvement. As a result of this work, interferon-targeted therapies have been developed and are presently being tested as treatment in SLE patients with the interferon associated gene expression profile[65,66].

Microarray studies in T1D have characterized peripheral blood gene expression signatures unique to both patients who test positive for islet-autoantibodies and patients with clinical diagnosis of T1D, each distinguishable from the signature of CTRL patients[67]. The signatures feature differential expression of genes associated with cellular metabolism and oxidative phosphorylation as well as interferon response genes, similar to SLE.

Gene expression profiling of peripheral blood has also further characterized the RA signature, delineating differential gene expression in patients based on their HLA profiles, treatment status and disease activity. Gene expression profiling of synovial tissue has also characterized gene expressions based on the type of lymphocyte infiltration seen microscopically[68]. These advances make it possible to use gene expression to further characterize disease and assess response to treatment.

Collectively, studies of microarray gene expression analysis in autoimmune diseases show that in addition to distinguishing patients with disease from CTRL subjects and their relatives, data gathered from peripheral blood gene expression microarray analysis can also inform on cellular functional deficits that may contribute to the pathogenesis of disease.

Peripheral Blood Gene Expression in Metabolic Conditions

We hypothesized that other common, complex diseases like T2D, CAD and their precursor, MetS, would also have distinct peripheral blood gene expression profiles that would enable distinction of these patients from CTRL and enhance our understanding of each disease or state. We sought to analyze their gene expression profiles in reference to CTRL, to each other and to an autoimmune disease.

T2D, CAD and MetS each feature a peripheral blood gene expression profile distinct from that of CTRL patients (Chapter II). However, the profiles of each of these states are indistinguishable from the others by support tree clustering. Comparison of T2D, CAD and MetS with the autoimmune disease RA showed that the three metabolic-related disorders more closely resemble each other than RA. Gene set analysis revealed common differential regulation of genes involved in the activation of the innate immune system with an emphasis on monocyte and macrophage regulation in CAD and involvement of the adaptive immune system, specifically T cells, in T2D.

Genetics and Gene Expression

A portion of the autoimmune peripheral blood gene expression profile can also be detected in first-degree relatives indicating that gene expression in this group of diseases is not entirely a function of an active disease process. Rather, a portion of the differential gene expression may be influenced by genetics or environmental exposures in the absence of disease. To test the hypothesis that genetics influence gene expression in autoimmune disease, investigators mapped the coding regions of differentially expressed

genes to the chromosomal level. Over 50% of the autoimmune signature genes mapped back to 1 of 15 chromosomal domains representing shared genetic loci that confer susceptibility to develop an autoimmune disease. Collectively, these loci are contained in less than 10% of the genome[56]. It is possible that inherited variations of the genome within these domains influence the expression of the genes in these regions, strengthening the argument for genetic inheritance influencing gene expression[56,69].

With the sequencing of the human genome, microarray technology expanded into genomics and arrays were built to assess genomic variation. Parallel measurement of genomic variation and gene expression in the same subject allowed for gene expression to be mapped back to the DNA coding sequence. In these studies, expression levels of mRNA transcripts are treated as "traits," mapped to their respective genomic coding regions, and identified as gene expression quantitative trait loci, or eQTLs[70]. These studies found that the majority of eQTLs are regulated by genetic variation in close proximity to the eQTL or in cis (on the same chromosome) rather than by distal genetic variants or in trans (on a different chromosome)[71].

Genetics

Applications of array technology to gene expression and genomics led to the first genome-wide association studies (GWAS). GWAS associate genomic variants with both physical traits, like height[72] and obesity[73] and with the incidence of common complex diseases like autoimmune diseases[74]. Genomic variation can be found at the single

nucleotide level and in larger nucleotide variations, such as those spanning > 1kb of genomic DNA.

Single Nucleotide Polymorphisms

Widespread genomic variation was first discovered in the form of single nucleotide polymorphisms (SNPs)[75]. A SNP is the change of a single base pair in a coding or noncoding sequence that tends to occur in association with other single nucleotide changes, forming haplotype blocks[76]. A haplotype block is a set of alleles that segregates together forming a unit of genetic inheritance. Haplotype blocks may also segregate together at a greater frequency than chance and these non-random associations are said to be in linkage disequilibrium within the genome. Haplotype blocks in linkage disequilibrium may each contain a variant in a gene or group of genes associated with a disease. The SNP or SNPs in that block then serve as measurable markers of the variance.

SNPs have been associated with common, complex diseases such as autoimmune diseases[26,77,78,79,80]. Many of the associations are shared among multiple autoimmune diseases, including association of SNPs in the HLA region with T1D, RA, MS and Crohn's disease[81,82,83]. The HLA region of chromosome 6 is the portion of the genome most robustly linked to autoimmune disease and was the only genetic region to be associated with progression to diabetes in a longitudinal study testing the associations of 37 SNPs in individuals at high risk for developing T1D[84].

Certain SNP associations confer a functional liability on the cell. For example, a SNP in the gene coding region of the intracellular lymphoid tyrosine phosphatase 22, PTPN22, is associated with T1D and RA among other autoimmune disorders[83,85]. In the

presence of the PTPN22 phosphatase polymorphism, the phosphatase shows reduced kinase binding, interfering with the activation of T cells[86]. This functional liability is directly related to a pathway, T cell activation, known to be involved in the pathogenesis of autoimmune disease.

Copy Number Variation

Copy number variations (CNVs) are amplifications or deletions in the genome that span more than 1kb of genomic DNA[87,88]. These alterations, which are widespread in nature, result in a deviation from the baseline 2 copies of each genomic segment per genome, to an amplification of 3 or more copies or a deletion to 1 or zero copies. One mechanism by which CNVs are generated is non-allelic homologous recombination (NAHR)[89]. NAHR is a normal meiotic process by which chromosomes align and recombine to generate the normal diversity seen in the human population. When identical regions of tandem repeats or regions of segmental duplication are present both upstream and downstream of a gene, it is possible for the chromosomes to misalign (Figure 1-3). The crossing over of misaligned chromosomes results in 2 copies of the gene being passed on to one daughter cell and zero copies of the gene to the other daughter cell. Assuming an identical event did not occur in the other haploid cell, the resultant diploid genomes would contain 3 copies and 1 copy of the gene, respectively.

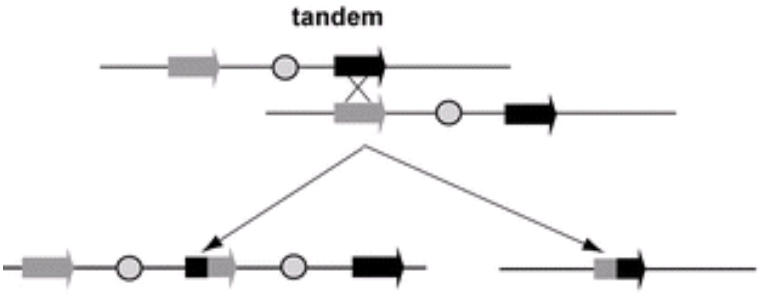

Figure 1-3. Schematic model of a molecular mechanism for meiotic NAHR between low copy repeats. Crossover between two paralogous low copy repeats (black and shaded rectangles) in direct orientation results in reciprocal duplication and deletion. These rearrangements lead to a duplication or deletion of a unique genomic segment (circle) flanked by the low copy repeats.
Taken from: Inoue, K and JR Lupski. "Molecular Mechanisms for Genomic Disorders." Annual Review of Genomics and Human Genetics. 2002. 3:199-242.

The influence of segmental duplications on the generation of CNVs is great. Segmental duplications can be defined as >1kb regions of the genome that share greater than 90% similarity with another region of the genome[90]. More than half of the nucleotides known to fall within regions of segmental duplication also comprise part of a known CNV and CNVs not associated with regions of segmental duplication tend to be, on average, less common in the population. Of note, regions of segmental duplication have a lower density of SNPs than the rest of the genome, indicating that these CNVs represent novel variation not previously assessed in SNP studies[91,92].

Meiotically generated CNVs, specifically gene amplifications, play an important role in evolution. Certain regions of the human genome recombine to generate CNVs at much higher rates than the rest of the genome. The olfactory receptor gene family is one of the largest in the human genome and the coding sequences are rich with segmental

duplications[93,94]. High degrees of sequence similarity found between consecutive olfactory receptor genes suggest that the initial duplication of genes through the process of copy number (CN) amplification is in part responsible for the size of this gene family.

CNVs may also be generated during mitotic processes such as non-homologous end joining, NHEJ[95]. NHEJ is an error-prone process of DNA repair following such events as replication fork stalling on the DNA template during DNA duplication[89]. This process is down regulated in meiosis but occurs frequently in mitosis. When NHEJ was chemically induced in clonal cell populations, one effect was formation of novel CN alterations in regions not characterized by tandem repeats or other known fragile breakpoints that may recombine during meiosis. These DNA repair induced variants are more likely to occur *in vivo* in continuously dividing cells, like lymphocytes for example. The result is somatic mosaicism, or the presence of a CNV in one tissue of an organism but not all[96].

CNVs can influence gene expression by dosage effects, or a proportional association between the number of copies of a gene and the number of transcripts. Other variants can interrupt coding sequences or alter gene regulation processes[97]. A genomic deletion can decrease the amount of protein through a dosage effect[98], and through transcriptional regulation effects. An example of the latter would be the deletion of a promoter region decreasing the production of a downstream gene transcript. Positional effects, or increasing and decreasing the distance between a gene and its promoter or repressor, is another mechanism by which CNVs influence gene expression. Finally, studies in mice have shown that a single CNV can affect expression of genes outside of the CNV and up to 1/2 megabase away[99,100]. For all of these reasons, it is difficult to

predict the potential functional outcome of a CNV without extensive knowledge of the genomic region housing the variant.

CNVs can be detected by fluorescence in-situ hybridization (FISH), quantitative PCR (qPCR), array comparative genomic hybridization (array-CGH) and SNP-based arrays[101]. FISH can be used to discover large CNVs and to confirm the presence or absence of a variant in a particular genome. FISH is fluorescent staining of chromosomes that requires the sequence of a region of interest be known. Thus, FISH is not a method by which to discover novel variants or associations. qPCR also requires a known target. While these accurate methods are ideal for investigating a known genomic region of interest or validation of a CN call determined by a different method, they do not lend themselves to genome-wide CN surveys.

Both array-CGH and SNP-based arrays have the ability to determine CN across the entire genome and consequently discover rare variants or novel associations[102]. Array-CGH functions much like a gene expression microarray. Clone-derived genomic sequences are robotically spotted onto a glass slide. Fluorescently labeled CTRL and test DNA are hybridized to the slide, the slide is scanned and intensity values are read for each sample at each spot. A ratio of the intensity values is used to determine CN across the genome. SNP-based genotyping arrays are built using matched and mismatched probes originally designed to detect SNPs; however, the same concept of probe design has been applied to regions without SNPs. Only one genomic DNA sample is hybridized to each array and the sample is first digested with restriction enzymes to improve the signal-to-noise ratio. Intensity values at the probes are interpreted using computer

algorithms. Array-CGH and SNP-based arrays have both been widely used to determine genome-wide CN and its potential associations with traits or diseases.

Intensity data from these arrays can be transformed into CN using any of a number of methods, including defined threshold intensity cut-offs and complex statistical algorithms like circular binary segmentation or Hidden Markov Models (HMM). These methods, however, are known to call widely different CN values for the same regions. Researchers often combine methods by using a "consensus copy number" from multiple algorithms, or validating CN by PCR. So far, there is not a standard method by which to determine accurate and reproducible CNs across the genome.

Copy Number Variation in the Human Genome

CNVs were first thought to be rare and, when present, pathogenic[103]. However, along with the ability to measure genome-wide CNV, we learned that these variants are in fact widespread in the genome and most often benign[88,104]. The International HapMap project was created to catalog global human genetic variation, beginning with analysis of >1 million SNPs in the genomes of 270 individuals from around the world- 30 parent-child trios from Nigeria, 30 parent-child trios of European descent living in Utah, 45 unrelated individuals of Japanese descent living in Tokyo and 45 unrelated Han Chinese individuals living in Beijing[105]. Redon, et al. determined CN in the HapMap 270 using a combination of SNP arrays and comparative genomic hybridization methods[106]. They discovered 1,447 CNV regions covering 12% of the genome with no portions of the genome seemingly exempt from this class of variant (Figure 1-4). 24% of these sequences were flanked by segmental duplications, more than expected by chance.

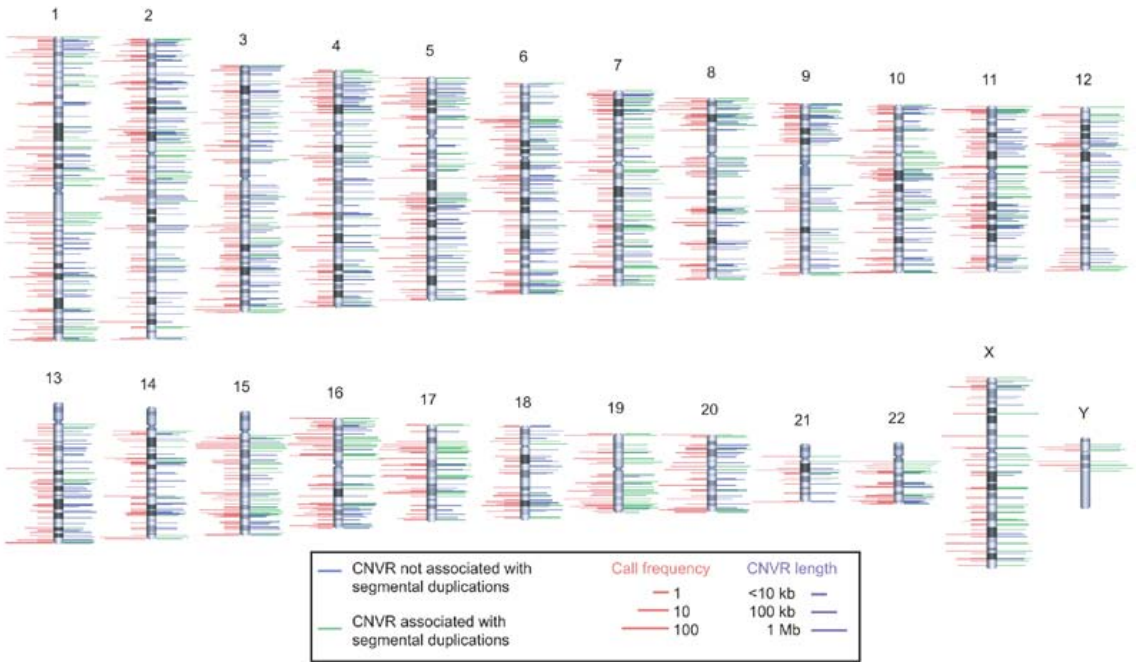

Figure 1-4. Genomic distributions of CNVRs.
The chromosomal locations of 1,447 CNVRs are indicated by lines to either side of ideograms. Green lines denote CNVRs associated with segmental duplications; blue lines denote CNVRs not associated with segmental duplications. The length of right-hand side lines represents the size of each CNVR. The length of left-hand side lines indicates the frequency that a CNVR is detected (minor call frequency among 270 HapMap samples). When both platforms identify a CNVR, the maximum call frequency of the two is shown. For clarity, the dynamic range of length and frequency are log transformed (see scale bars). All data can be viewed at the Database of Genomic Variants (http://projects.tcag.ca/variation/).
Taken from: Redon, R. et al. "Global variation in copy number in the human genome." Nature. 2006. 444: 444-454.

McCarroll, et al. discovered 3,048 CNV regions in the HapMap population using a SNP-based microarray modified to include nucleotide probes specific for CN. 1,320 of these CNVs segregated at allele frequencies greater than 1% and were thus denoted as copy number polymorphisms (CNPs). The CNPs span approximately 5% of the genome[107]. The percent of the genome that differed from one subject to the next was less than 0.5%, with more than 90% of these differences classified as CNPs. Family trios

within the HapMap also showed that inheritance, not de novo generation, is responsible for the majority of CNVs.

Additional studies have varied in their estimation of genome-wide CNV. Estimates of the number of distinct regions of CNV present in the genome range from >1,000 to >4,000, spanning 2.86% to 25% of the genome[90,106,107,108,109,110]. The average length of a CNV in these studies is 200-300kb[109,110]. Variables like sample size and resolution, related to the different experimental methods and statistical models used to determine CN in these studies account for the lack of consensus regarding the amount of the genome that can be variant by CN. However, if all studies are considered together, the 38,406 non-independent CNVs deposited to the Database of Genomic Variants (DGV), cover 29.74% of the genome. For comparison, less than 1% of the genome is affected by the 14 million non-independent DGV-deposited SNPs[111].

CNVs are known to be associated with a number of diseases, both Mendelian disorders (diseases caused by the mutation of one gene) and common complex diseases. One of the first reported CNV-disease associations was the duplication of the gene PMP22 as a direct cause of Charcot-Marie Tooth disease[112]. Increased copies of PMP22 correspond with increased gene transcripts through a dosage effect resulting in phenotypic disease.

A CNV in the *FCGR3A* gene, encoding the Fc-gamma receptor, is associated with SLE[113]. This protein binds the Fc portion of the antibody IgG and variation impacts protein expression in a dose-dependent manner[114]. Functionally, decreased levels of FcγRIIIa impact the ability of NK cells to carry out antibody-dependent cytotoxicity. CNVs in FCGR3B have also been associated with SLE[115,116].

Additionally, CNV in the gene CCL3L1 is associated with susceptibility to infection with HIV[117] and CNVs are also associated with a number of neurological, neuropsychological and behavioral disorders like epilepsy, Parkinson, schizophrenia and autism[118,119,120,121,122]. Deletions in the LCE genes have been associated with psoriasis and deletions in the complement factor H genes protect from age-related macular degeneration[123,124,125].

Copy Number Variation in Type 1 Diabetes

In spite of the numerous SNP associations made in autoimmune diseases, SNPs have so far failed to fully explain disease pathogenesis or account for genetic inheritance. The second phase of the human HapMap project catalogued over 3.1 million human SNPs genotyped in the first 270 HapMap individuals[126]. A genome-wide association study for SNPs associated with any 1 of 7 common diseases, including T1D, failed to identify many new SNP associations and verified that the majority of SNP associations confer very small risks for disease. The authors further determined that known SNPs do not account for all of the familiality of the 7 diseases studied and concluded that there must be another source of genetic variation responsible for heritability[26].

We hypothesized that CNVs would be found enriched and depleted in patients with T1D. To test this hypothesis, CNVs were analyzed using the Affymetrix SNP 6.0 array. This array contains nearly 1 million probes designed to detect all known SNPs and an additional 1 million probes specifically designed to assess genome-wide CN. We first addressed the problem that different calling algorithms produced different CN calls based on the same data by utilizing qPCR to inform validity of the calls (Chapter III). This

process included an evaluation of three commonly used algorithms and resulted in a reliable and reproducible method by which to call CN in our samples.

With these established methods, CNVs were detected to be both enriched and depleted in a cohort of patients with T1D and an independent cohort of monozygotic twins discordant for T1D (Chapter IV). Many of the CNV regions were similarly enriched or depleted in RA and MS, indicating that these genomic markers may encode regions important in the development of autoimmunity and autoimmune disease.

CHAPTER II

PERIPHERAL BLOOD GENE EXPRESSION PROFILES IN METABOLIC SYNDROME, CORONARY ARTERY DISEASE AND TYPE 2 DIABETES

Abstract

To determine if peripheral blood from individuals with inflammatory metabolic disorders has a distinct gene expression profile distinguishable from CTRL individuals, we performed comprehensive analysis of transcript levels in peripheral blood from patients with coronary artery disease, type 2 diabetes and their precursor state, metabolic syndrome. We compared these gene expression profiles to those of CTRL subjects and patients with a classic autoimmune disease, rheumatoid arthritis. The gene expression profile of each metabolic state was distinguishable from that of CTRLs and the 3 states showed greater correlation with each other than with rheumatoid arthritis. Of note, subjects with metabolic syndrome, coronary artery disease or type 2 diabetes over-expressed genes and gene sets involved in the innate immune response. Genes involved in activation of the pro-inflammatory transcription factor, NF-κB, were also over-expressed in coronary artery disease. Many genes differentially expressed in type 2 diabetes regulate T cell activation and signaling. RT-PCR analysis of a subset of genes identified in this pathway analysis determined quantitative differences in the disease versus CTRL comparison but additional overlap of significance among the metabolic disorders, greater than suggested by array analysis. Additionally, many of these genes

correspond to genes differentially expressed in both mouse models or other human studies of insulin resistance and obesity. Taken together, these data demonstrate that the peripheral blood from individuals with metabolic disorders display both overlapping and non-overlapping patterns of gene expression indicative of unique, underlying immune processes.

Introduction

Type 2 diabetes (T2D) is a metabolic disorder of peripheral insulin resistance resulting in hyperglycemia and ultimately decreased insulin secretion from the pancreas. Risk factors for T2D include obesity, physical inactivity and family history[28]. Diabetes currently affects 6.3% of the United States population and approximately 90% of these cases are non-insulin dependent, or T2D[127]. Coronary artery disease (CAD) results from atherosclerotic plaque development in coronary arteries. These fibrous, fatty deposits can ultimately block the flow of blood resulting in angina and/or myocardial infarction. Hyperlipidemia predisposes to the development of these plaques, making obesity and physical inactivity also risk factors for CAD[30]. The prevalence of CAD in the United States is 4.1%[128]. Metabolic syndrome (MetS) is a precursor state to both T2D and CAD[2,3,32,33,34,35,36,37,38]. The International Diabetes Federation (IDF) defines this pre-disease state as central obesity plus any 2 of the following 4 characteristics: hypertriglyceridemia, low high-density lipoprotein (HDL) cholesterol, hypertension or raised fasting plasma glucose. The prevalence of MetS in the United States is as high as 39% using the IDF criteria. Diagnosis of MetS confers a 1.5-2.6 relative risk of

developing CAD and a 3.5-7.5 relative risk of developing T2D. Additionally, the Framingham study determined that a portion of these relative risks persist even in the absence of obesity. This trio of disorders poses a significant threat to public health in the United States.

Inflammatory processes are involved in the pathogeneses of T2D and CAD. Visceral adipose tissue, present in abundance in many patients with T2D, produces inflammatory cytokines like IL-6 and TNF-α that are known to aid in the impairment of insulin signaling in adipocytes. These cytokines can activate a systemic immune response and attract inflammatory cells, like lymphocytes, to infiltrate visceral adipose tissue [129]. In the case of CAD, the lesion is not visceral adipose tissue, but rather fatty deposits in the vasculature. These deposits contain fat-laden macrophages and immunoreactive T-cells[30].

Gene expression profiling of blood or tissue samples is one way to assess cellular changes due to cell differentiation and aging[130,131], disease pathogenesis[132,133,134] or pharmacological response[135,136]. One example of this is tumor typing; gene expression signatures are presently used to classify tumor types in breast cancer biopsies. This method can also be used to assess changes in peripheral whole blood of patients with common, complex diseases[55,137,138]. Individuals with autoimmune diseases [type 1 diabetes, multiple sclerosis, systemic lupus erythematosus and rheumatoid arthritis (RA)] display unique gene expression signatures in peripheral whole blood. Portions of these signatures are expressed in first degree unaffected relatives[57]; however, disease-specific signatures are also found in peripheral blood and are sufficient to distinguish individuals with disease from CTRL individuals[58]. Moreover, peripheral blood gene expression

profiling can give insight into disease processes and suggest specific functional defects in cells. For example, peripheral blood gene expression in patients with RA contains low transcript levels of the tumor suppressor protein, p53. Consequently, T cells from patients with RA are resistant to gamma-radiation induced apoptosis, a p53 dependent pathway[59]. Gene expression profiling may also aid in diagnosing patients who have these often hard to diagnose diseases; therefore, analysis of peripheral blood gene expression already represents one way to assess immune system changes, predict related cellular defects and diagnose patients with immune-related disease in a minimally invasive way.

T2D, CAD and their precursor, MetS are not autoimmune diseases but feature inflammation as a possible pathogenic component. The purpose of our studies was to assess if these inflammatory diseases also display unique peripheral blood gene expression profiles and if so, what do the profiles indicate about the interrelatedness of MetS, CAD and T2D. To address this question, we compared profiles of each disease state to control (CTRL) subjects, to an autoimmune disease, RA, and to each other. We found that MetS, CAD and T2D have unique gene expression profiles that distinguish each cohort from CTRL subjects. Further, profiles of MetS, CAD and T2D are more similar to each other than to RA. These profiles feature a common component of activation of the innate immune response, reflected by increased expression of complement factor related proteins and acute phase proteins, for example. Over-expression of genes involved in activation of the pro-inflammatory transcription factor, NF-κB, was a dominant theme in the CAD expression signature, whereas the T2D profile included increased expression of genes encoding proteins that regulate activation and

signaling in T cells. Additionally, many of these genes correspond to genes differentially expressed in both mouse models or other human studies of insulin resistance and obesity.

Materials and Methods

Patient Recruitment

Rheumatoid arthritis is defined by the American College of Rheumatology Criteria. Patients displayed four or more of the following symptoms for greater than 6 months: morning stiffness, swelling in 3 or more joints, swelling of finger and/or wrist joints, symmetric swelling, rheumatoid nodules, positive rheumatoid factor, or radiographic erosions in the hand and/or wrist[139]. *Metabolic syndrome* is defined by the International Federation of Diabetes as central obesity plus any 2 of the following 4 characteristics: hypertriglyceridemia, low HDL cholesterol, hypertension or raised fasting plasma glucose[40]. C*oronary artery disease* was diagnosed in each patient using imaging techniques to detect flow-limiting coronary artery stenoses[140]. Three of the 6 patients with coronary artery disease participating in this study were post coronary artery bypass graft or myocardial infarction. All patients in this cohort are also being treated for systemic hypertension. *Diabetes* is defined by the WHO criteria of classic symptoms of diabetes (polydipsia, polyuria, polyphagia and weight loss) and a plasma glucose >200 mg/dl, a fasting plasma glucose of >126 mg/dl or a 2 h plasma glucose during an oral glucose tolerance test of >200 mg/dl. Type 1 diabetes is differentiated from T2D by a number of clinical criteria including history, clinical presentation and laboratory findings. Type 2 diabetics are more likely to have a high body mass index and less likely to need

insulin in restoring normal plasma glucose levels[32]. *Control* patients have not ever received any of the previous diagnoses, have not been diagnosed with any autoimmune or other chronic disease, and are not currently taking medication for any illness or condition. The study was approved by the Institutional Review Board of Vanderbilt University and all subjects provided written informed consent.

Microarray Gene Expression Experiments

Peripheral whole blood was drawn directly into PreAnalytiX PAXgene tubes (VWR, West Chester, PA). RNA was isolated using the PreAnalytiX protocol "Manual Purification of Total RNA from Human Whole Blood Collected into PAXgene Blood RNA Tubes." Amplified CTRL and sample RNA was coupled to Cy3 or Cy5 dyes (GE Healthcare, Piscataway, NJ), respectively, using the Vanderbilt Functional Genomics Shared Resource (FGSR) coupling protocol, found at [array.mc.vanderbilt.edu]. The reverse transcription reaction used 6 μg of Oligo dT and the superscript III reverse transcriptase (Invitrogen, Carlsbad, CA). Labeled cDNA was purified using the Qiagen QiaQuick PCR purification kit and resuspended in 2X hybridization buffer (50% formamide, 10X SSC and 0.2% SDS) and 1 μl polyA RNA. Labeled, resuspended cDNA was heated to 100ºC for 2 min and hybridized to the Human Exonic Evidence-Based Oligonucleotide (HEEBO) array at 42ºC for 16 h in a heating oven. The HEEBO slide was designed by the Stanford Functional Genomics Facility (microarray.org). Oligo probes are commercially available (Invitrogen, Carlsbad, CA) and the slides were printed by Microarrays Inc (Hudson Alpha Institute, Huntsville, AL). We washed and dried the slides per the FGSR protocol and scanned them into the GenePix Pro4.1 Software using a

400B scanner (Axon Instruments, Union City, CA). We analyzed intensity data using GenePix software in combination with The Institute for Genomic Research's TM4: Microarray Suite programs[43].

Microarray Data Analysis

The Institute for Genomic Research's Multi-Experiment Viewer was used to visualize intensity data. We used Significance Analysis of Microarray to determine a group of significantly under- and over-expressed genes in the comparisons of each disease group versus CTRL. The median number of falsely significant genes was set to ≤2. Following this analysis, the bootstrap statistical method created support trees showing hierarchical clustering for the 4 comparisons of each disease versus CTRL based on 1,000 permutations. For statistical analysis of gene sets, we normalized microarray data using the print-tip lowess normalization algorithm as implemented in the Bioconductor package *marray*[141]. We used maximum expression levels from multiple probe sets corresponding to the same gene to represent the gene expression level. To ensure reliable gene expression estimates, we included genes with intensity values for more than 6 CTRL samples and more than 3 samples for each of the other groups. There were 14,558 genes left after this step.

To identify groups of functionally related genes differentially expressed for different patient groups, we conducted gene set analysis using the mixed effects models approach[142,143]. Gene sets used in these analyses were derived from the controlled vocabulary of the Gene Ontology (GO) project, http://www.broad.mit.edu/gsea/msigdb/index.jsp. For each gene set, the mixed models

included gene expression levels as outcome, group (disease group vs. CTRL group) as the fixed effect and batches as the random effects. In addition, we included random effects based on eigenvectors of gene-gene correlation matrix to account for correlation patterns of the genes[143]. Because we examined many gene sets, to control for the rate of false positive findings by chance, we adjusted nominal *p*-values using the method of false discovery rate[144]. To study the relations between T2D, MetS, CAD and RA, we estimated pairwise Spearman correlation coefficients for these disease groups based on nominal pathway *p*-values from comparing each disease group versus CTRL. We used Cytoscape software[145] to visualize these associations.

The data discussed in this chapter have been deposited in NCBI's Gene Expression Omnibus[146] and are accessible through GEO Series accession number GSE23561 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23561).

RT-PCR

Gene expression was determined by RT-PCR using a TaqMan Low Density Array (TLDA). Fold change expression levels were determined by the ΔΔCt method, comparing expression of test gene to an average of two independent measurements of GAPDH, and then comparing the disease cohort versus CTRL. Significance was determined using a t-test on the ΔCt raw values.

Results

Peripheral blood gene expression profiling using microarrays has been shown sufficient to distinguish between phenotypically distinct cohorts of patients[55,137,138]. We sought to determine if subjects with MetS, CAD or T2D also possessed a gene expression signature in blood sufficient to distinguish these subjects from CTRL subjects and, if so, did this signature bear any resemblance to the signature of an autoimmune disease, RA. To do so, we recruited subjects with MetS, CAD and T2D (n=6, n=6, n=8, respectively), 6 subjects with RA, and 9 subjects who had never been diagnosed with a chronic illness, and were not presently taking medications for any diagnosed state, to serve as the CTRL cohort.

We analyzed all 35 peripheral blood samples for gene expression using the human exonic evidence-based oligonucleotide (HEEBO) array. Next, we normalized the data to a sum total intensity of 10,000, giving an average intensity per oligonucleotide probe of 0.2. Those genes with an average intensity of greater than 0.2 were used as the data points for clustering analysis. The intensity values of the filtered set of genes for each array were inputted into The Institute for Genomic Research's multi-experiment viewer. We compared the RA, MetS, CAD and T2D groups individually to the CTRL cohort using a supervised significance analysis for microarray function. Each list of significantly differentially expressed genes was used to run a bootstrap hierarchical clustering to determine the similarity of the patient samples to each other within each disease and their similarity to CTRL (Figure 2-1). A black bar, representing 100% support, separates two main branches neatly clustering the RA cohort away from the CTRL cohort based on

gene expression. For the T2D cohort, two T2D patients clustered with the CTRL group and conversely, the same 2 CTRL patients clustered on a branch with the MetS and CAD groups in their individual comparisons with CTRL. These separations indicate that the majority of persons with RA, MetS, CAD or T2D are more like each other than the CTRL. Hierarchical clusters confirm that expression of genes in peripheral whole blood is sufficient to distinguish between the autoimmune disease, RA, and CTRL, as well as the inflammatory metabolic states of MetS, CAD and T2D and CTRL. Further similarities and differences can be seen amongst the disease affected subjects. In the RA group at least one further branch with 100% support was seen, indicating that gene expression is not entirely homogenous within this group.

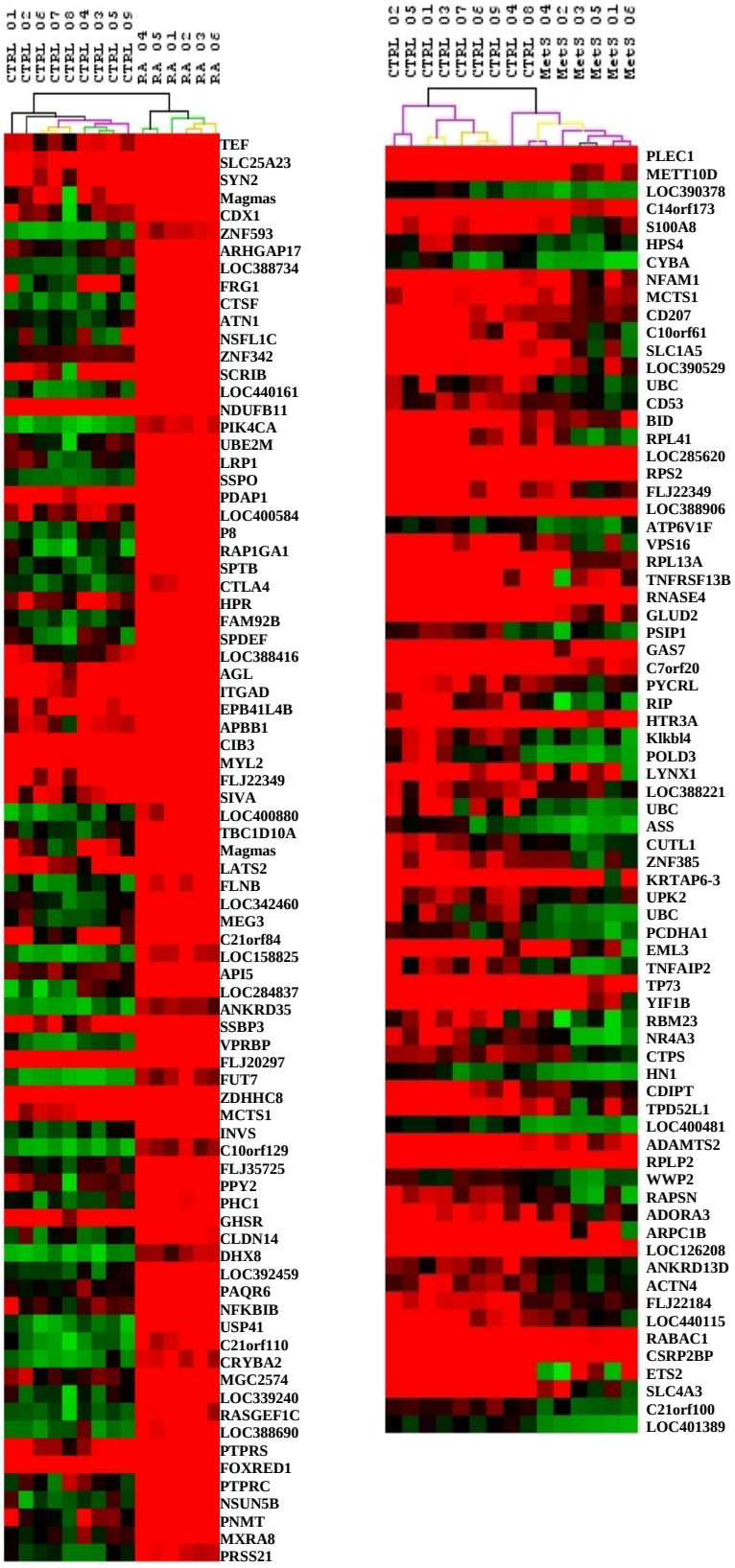

Figure 2-1. Hierarchical clustering of individual disease cohorts versus CTRL.
(continued on next page)

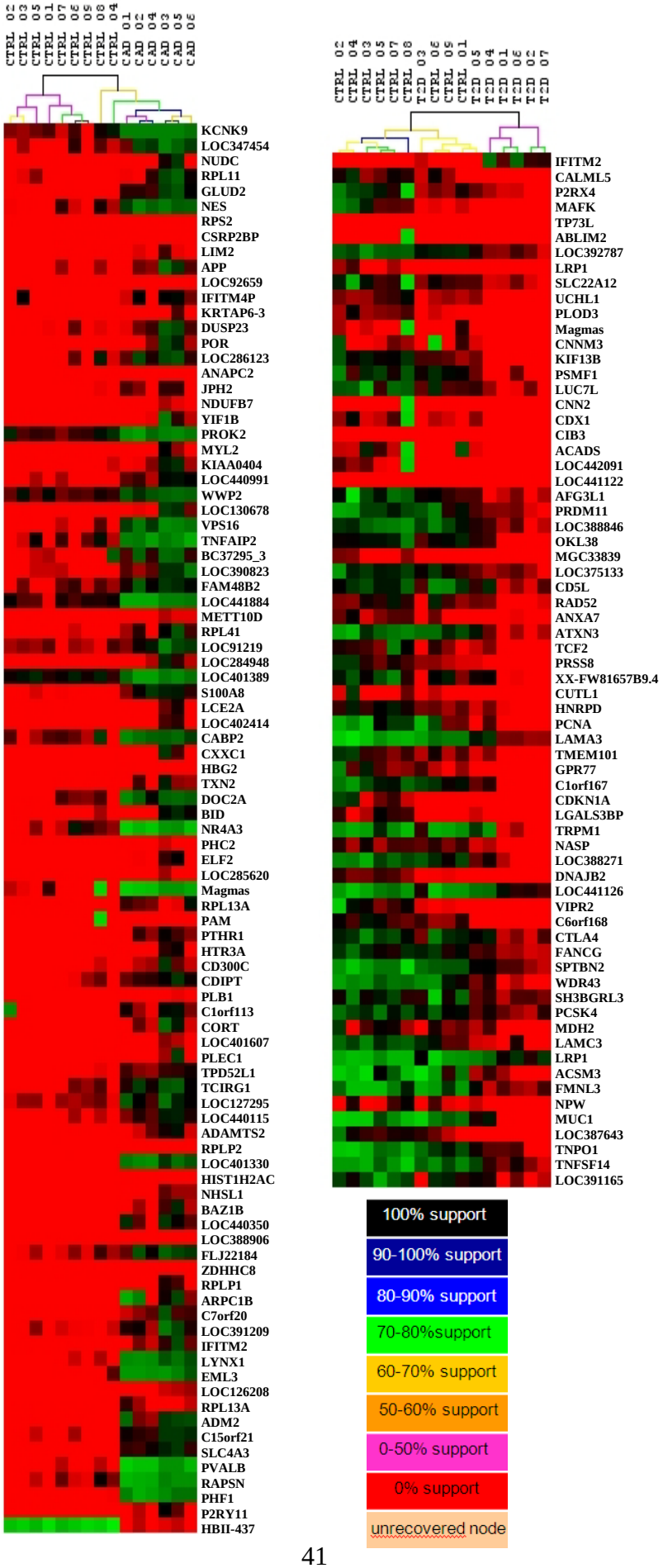
CTRL 02
CTRL 03
CTRL 05
CTRL 01
CTRL 07
CTRL 06
CTRL 08
CTRL 04
CAD 01
CAD 02
CAD 04
CAD 03
CAD 05
CAD 06
KCNK9
LOC347454
NUDC
RPL11
GLUD2
NES
RPS2
CSRP2BP
LIM2
APP
LOC92659
IFITM4P
KRTAP6-3
DUSP23
POR
LOC286123
ANAPC2
JPH2
NDUFB7
YIF1B
PROK2
MYL2
KIAA0404
LOC440991
WWP2
LOC130678
VPS16
TNFAIP2
BC37295_3
LOC390823
FAM48B2
LOC441884
METT10D
RPL41
LOC91219
LOC284948
LOC401389
S100A8
LCE2A
LOC402414
CABP2
CXXC1
HBG2
TXN2
DOC2A
BID
NR4A3
PHC2
ELF2
LOC285620
Magmas
RPL13A
PAM
PTHR1
HTR3A
CD300C
CDIPT
PLB1
C1orf113
CORT
LOC401607
PLEC1
TPD52L1
TCIRG1
LOC127295
LOC440115
ADAMTS2
RPLP2
LOC401330
HIST1H2AC
NHSL1
BAZ1B
LOC440350
LOC388906
FLJ22184
ZDHHC8
RPLP1
ARPC1B
C7orf20
LOC391209
IFITM2
LYNX1
EML3
LOC126208
RPL13A
ADM2
C15orf21
SLC4A3
PVALB
RAPSN
PHF1
P2RY11
HBII-437
CTRL 02
CTRL 04
CTRL 03
CTRL 05
CTRL 07
CTRL 08
T2D 03
CTRL 06
CTRL 09
CTRL 01
T2D 05
T2D 04
T2D 01
T2D 06
T2D 02
T2D 07
IFITM2
CALML5
P2RX4
MAFK
TP73L
ABLIM2
LOC392787
LRP1
SLC22A12
UCHL1
PLOD3
Magmas
CNNM3
KIF13B
PSMF1
LUC7L
CNN2
CDX1
CIB3
ACADS
LOC442091
LOC441122
AFG3L1
PRDM11
LOC388846
OKL38
MGC33839
LOC375133
CD5L
RAD52
ANXA7
ATXN3
TCF2
PRSS8
XX-FW81657B9.4
CUTL1
HNRPD
PCNA
LAMA3
TMEM101
GPR77
C1orf167
CDKN1A
LGALS3BP
TRPM1
NASP
LOC388271
DNAJB2
LOC441126
VIPR2
C6orf168
CTLA4
FANCG
SPTBN2
WDR43
SH3BGRL3
PCSK4
MDH2
LAMC3
LRP1
ACSM3
FMNL3
NPW
MUC1
LOC387643
TNPO1
TNFSF14
LOC391165
100% support
90-100% support
80-90% support
70-80%support
60-70% support
50-60% support
0-50% support
0% support
unrecovered node

Figure 2-1. Hierarchical clustering of individual disease cohorts versus CTRL.
To determine if the gene expression profiles of the disease cohorts were distinguishable from that of the 9 CTRL patients, normalized intensity data points from oligos with an average intensity of ≥0.20 (average array intensity) were inputted into The Institute for Genomic Research's Multi-Experiment Viewer. For each comparison, gene intensity averages were calulcated and those≥0.20 were selected as input in each comparison. The CTRL v RA input was 4,969 gene and gene splice data points; CTRL v MetS input was 4,225 data points; CTRL v CAD input contained 4,271 data points and the CTRL v T2D comparison featured an input of 4,983 data points. Supervised significance analysis of microarray, with a median number of falsely significant genes set to≤2 , yielded lists of significant genes in each comparison, visible to the right of each heat map. Green indicates decreased expression while red denotes increased expression. These lists were inputted into a bootstrap analysis resulting in the hierarchical clustering trees shown above each map. Statistical support for each branch of the tree is shown by color, legend in the bottom right corner. CTRL= control, RA= rheumatoid arthritis, T2D= type 2 diabetes, MetS= metabolic syndrome and CAD= coronary artery disease.

Additionally, when all samples were analyzed together, the support tree indicates that 8/9 CTRL subjects cluster, with 100% support, on one branch with 2 T2D subjects, one of those being patient T2D 03 who previously clustered on the CTRL branch in the CTRL:T2D analysis (Figure 2-2). The second branch features just one CTRL patient and the remainder of the patients in the disease cohorts. Five of the 6 RA patients cluster together on this branch, indicating that the RA signature is more like that of the metabolic diseases than CTRL; however, the RA patients are more like each other than the metabolic cohort patients. Furthermore, the remaining metabolic disease patients did not cluster in any particular pattern suggesting similarity amongst the MetS, CAD and T2D peripheral blood gene expression profiles. Taken together, this analysis demonstrates that subjects with MetS, CAD or T2D each possess a common gene expression signature in blood sufficient to distinguish them from CTRL and that these signatures may have overlapping components.

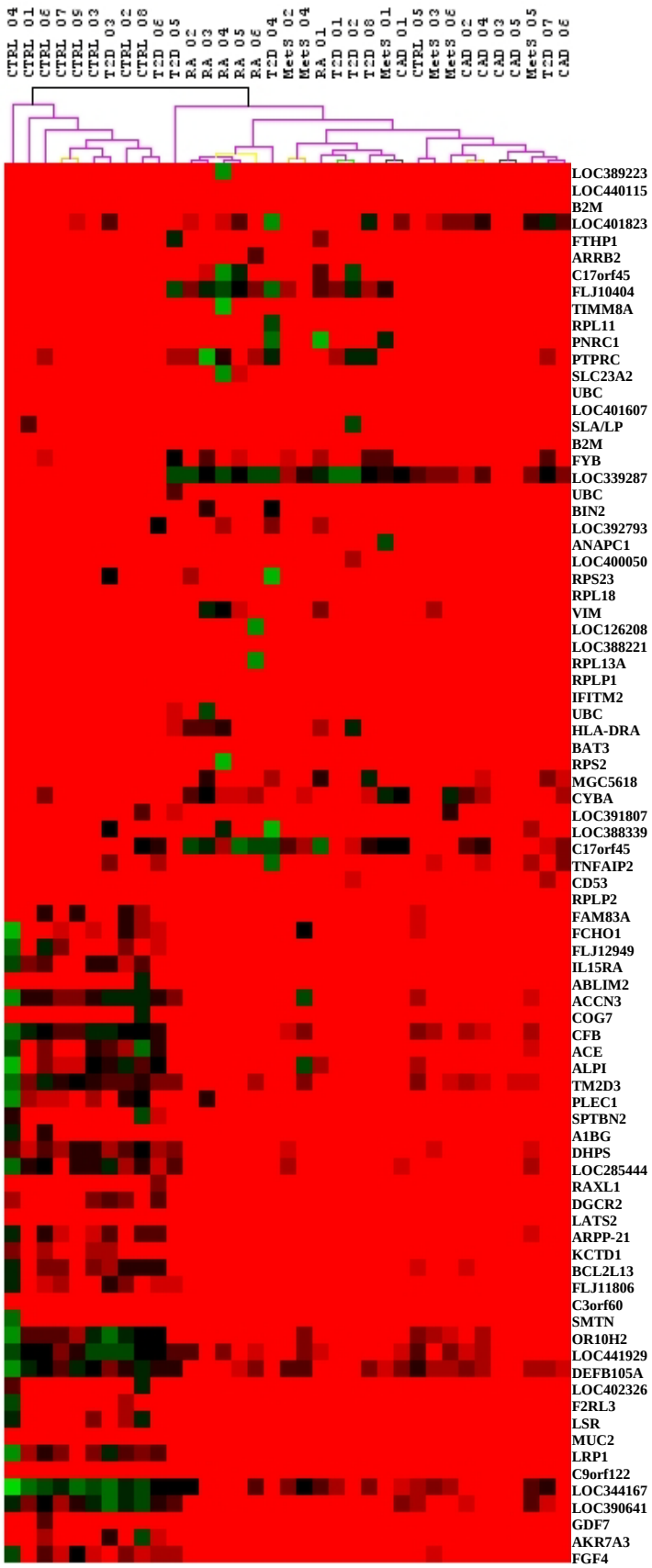
CTRL 04
CTRL 01
CTRL 06
CTRL 07
CTRL 05
CTRL 03
T2D 03
CTRL 02
CTRL 08
T2D 06
T2D 05
RA 02
RA 03
RA 04
RA 05
RA 06
T2D 04
MetS 02
MetS 04
RA 01
T2D 01
T2D 02
T2D 08
MetS 01
CAD 01
CTRL 05
MetS 03
MetS 06
CAD 02
CAD 04
CAD 03
CAD 05
MetS 05
T2D 07
CAD 06
LOC389223
LOC440115
B2M
LOC401823
FTHP1
ARRB2
C17orf45
FLJ10404
TIMM8A
RPL11
PNRC1
PTPRC
SLC23A2
UBC
LOC401607
SLA/LP
B2M
FYB
LOC339287
UBC
BIN2
LOC392793
ANAPC1
LOC400050
RPS23
RPL18
VIM
LOC126208
LOC388221
RPL13A
RPLP1
IFITM2
UBC
HLA-DRA
BAT3
RPS2
MGC5618
CYBA
LOC391807
LOC388339
C17orf45
TNFAIP2
CD53
RPLP2
FAM83A
FCHO1
FLJ12949
IL15RA
ABLIM2
ACCN3
COG7
CFB
ACE
ALPI
TM2D3
PLEC1
SPTBN2
A1BG
DHPS
LOC285444
RAXL1
DGCR2
LATS2
ARPP-21
KCTD1
BCL2L13
FLJ11806
C3orf60
SMTN
OR10H2
LOC441929
DEFB105A
LOC402326
F2RL3
LSR
MUC2
LRP1
C9orf122
LOC344167
LOC390641
GDF7
AKR7A3
FGF4

Figure 2-2. Hierarchical clustering of all disease cohorts versus CTRL.
To determine the similarity and difference of the profiles of each disease cohort to each other in the presence of CTRL, normalized intensity data points from all gene and gene splice oligos were inputted into The Institute for Genomic Research's Multi-Experiment Viewer (40,538 data points). Supervised significance analysis of microarray, with a median number of falsely significant genes set to 1.0, yielded a list of significant genes shown to the right of the heat map. This list was inputted into a bootstrap analysis resulting in the hierarchical clustering tree shown above the heat map. Green indicates decreased expression while red denotes increased expression. CTRL= control, RA= rheumatoid arthritis, T2D= type 2 diabetes, MetS= metabolic syndrome and CAD= coronary artery disease.

One possible source of differential gene expression in leukocytes is an alteration in the underlying genetic code[71]. Extensive genome wide analyses have been performed in RA, CAD and T2D revealing a number of single nucleotide polymorphisms (SNPs) associated with each individual disease. We probed our expression dataset to determine if genes associated with these SNPs showed differential expression in peripheral blood of subjects with disease versus CTRL subjects in any of our cohorts. A list of SNPs associated with RA, CAD or T2D was populated from The National Human Genome Research Institute[147] and a recent pathway based SNP analysis by Torkamani, *et al.*[148]. For the SNPs present in gene coding regions, we calculated expression levels of the encoded gene as an average for the RA, CAD and T2D groups. Each set of genes was analyzed for expression in disease groups and we found a number of correlations between a SNP, its encoded gene and differential expression of that gene. Eight genes with a disease-associated SNP were differentially expressed in the corresponding disease group- *CD244*, *IL2RA*, *PRKCA*, *SLC22A4* and *TRAF1* in RA, and *ADAMTS9*, *ANXA11* and *KCNQ1* in T2D (Table 2-1). *IL2RA* and *TRAF1*, genes identified by SNP studies in RA, were also differentially expressed in T2D and the T2D SNP-identified genes *ADAMTS9* and *KCNQ1* were differentially expressed in RA. While SNPs are known to influence

gene expression, we only found associations in RA and T2D, not CAD. Altered gene expression was not confined to just one disease state; differential expression of certain genes was shared between RA and T2D.

Table 2-1. SNPs associated with RA and T2D show differential gene expression

		RA		T2D	
RA SNP	**Gene**	***p*[a]**	**FC[b]**	***p***	**FC**
Rs6682654	*CD244*	0.009	5.57	ns	
Rs2104286	*IL2RA*	0.002	4.30	0.028	2.52
-[c]	*PRKCA*	0.001	2.88	ns	
-	*SLC22A4*	0.044	0.32	ns	
Rs3761847	*TRAF1*	0.026	0.34	0.010	0.29
T2D SNP					
Rs4607103	*ADAMTS9*	0.003	11.69	0.030	4.65
Rs2789686	*ANXA11*	ns		0.028	0.016
Rs2237892	*KCNQ1*	0.049	0.59	0.020	0.58

[a]*p*= derived from Mixed Effects Model, RA or T2D relative to CTRL
[b]FC= fold change, average of RA or T2D cohort relative to average of CTRL
[c]= identified via pathway-based analysis in Torkamani, *et al.*
ns= not significant

Because hierarchical clustering demonstrated differences in gene expression profiles of each metabolic disorder cohort versus CTRL and potential overlap amongst the signatures of the metabolic states, we further analyzed the relationships of gene expression within and amongst the 4 disease or pre-disease states in the context of gene sets. A gene set is defined as a group of genes with a common purpose, derived from the Gene Ontology project[44]. For further information on gene sets, normalization, and calculations, see the Methods section. Complete analysis with *p*-values for each gene set as well as the *p*-value and fold change for individual genes considered in each gene set comparison are also available (Supplemental Tables 2-1 and 2-2). Gene set analysis

showed that genes driving the differential expression in MetS, CAD and T2D are associated with overlapping activation of the innate immune response, activation of the pro-inflammatory transcription factor NF-κB in CAD, and over-expression of genes involved in T cell activation and signaling in T2D.

Rheumatoid Arthritis

Rheumatoid Arthritis is an autoimmune disease characterized by systemic inflammation that extends into and damages peripheral joints[149]. Patients with RA have robust and distinguishable gene expression profiles in peripheral whole blood[55]. This finding was repeated using the HEEBO slide as the array format. Our analysis identified 5 gene sets of particular significance (Table 2-2). *BIRC4* is over-expressed in gene set 110, Cell Development, and is involved in activation of the transcription factor NF-κB. NF-κB regulates expression of many pro-inflammatory genes. Immune System Process, gene set 271, includes over-expression of *LAT2* and *NFAM1*, genes involved in B cell signaling and development. Additional genes, *BAT1*, *LIG4* and *ILF2*, are expressed in lymphocytes and differentially expressed in gene set 435. *BAT1* is an HLA-associated transcript mutated in patients with RA. *LIG4* encodes a protein essential for V(D)J recombination and non-homologous end joining as part of DNA repair. *ILF2* is involved in T cell expression of IL-2, a potent stimulator of proliferation of lymphocytes. The IL2-receptor alpha, *IL2RA*, is also over-expressed in this cohort and is found in gene set 753, Signal Transduction. Differential expression of genes involved in activation, maturation and signaling of lymphocytes is in agreement with the gene expression profile of RA seen previously[55].

Table 2-2. Differentially expressed gene sets

Gene Set	Gene Set Name	*p*-value
RA v CTRL		
110	Cell Development	0.0045
271	Immune System Process	0.0116
435	Nucleobase Nucleoside Nucleotide and Nucleic Acid Metabolic Process	2.13E-08
706	Response to External Stimulus	0.0078
753	Signal Transduction	1.03E-11
MetS v CTRL		
13	Acute Inflammatory Response	0.048
316	Lymphocyte Differentiation	0.004
407	Negative Regulation of Signal Transduction	0.051
615	Regulation of Developmental Process	0.023
CAD v CTRL		
412	Negative Regulation of Transferase Activity	0.014
499	Positive Regulation of Immune Response	0.020
636	Regulation of I KappaB Kinase NF KappaB Cascade	0.051
T2D v CTRL		
104	Cell Cell Signaling	0.0048
117	Cell Proliferation Go 0008283	0.002
271	Immune System Process	1.7E-06
435	Nucleobase Nucleoside Nucleotide and Nucleic Acid Metabolic Process	9.7E-28
753	Signal Transduction	4.8E-13
CAD v MetS		
372	Negative Regulation of Biological Process	5.7E-04
482	Positive Regulation of Cellular Process	0.008
682	Regulation of Transcription	0.019
753	Signal Transduction	0.009
T2D v MetS		
271	Immune System Process	0.043
478	Positive Regulation of Caspase Activity	0.033
596	Regulation of Cellular Metabolic Process	0.052
104	Cell Cell Signaling	0.030

Other genes significantly differentially expressed in this gene set included the IL9-receptor, *IL9R*, which supports IL-2 and IL-4 independent T cell growth, and *MAP2K6*, which activates p38 MAP kinase in response to inflammatory cytokines. *LILRB4* was significantly under-expressed as part of the Signal Transduction gene set. This gene is an immune-cell receptor for MHC-I that transduces a signal to inhibit the immune response; increased expression of LILRB4 on antigen presenting cells renders the cells tolerant, therefore decreased expression might allow for increased autoreactivity[150]. Finally, in gene set 706, Response to External Stimulus, *CHST2*, encoding a protein expressed by vascular endothelium to attract lymphocytes, and *F11R*, encoding another protein expressed by vascular endothelium and involved in leukocyte transmigration, were over-expressed. Over-expression of genes encoding proteins with key roles in lymphocyte activation and growth could influence activation and expansion of self-reactive lymphocytes believed to cause joint destruction in individuals with RA.

Metabolic Syndrome

The triad of MetS, CAD, and T2D are typically considered metabolic, not immune, diseases although aspects of each involve inflammation, sometimes systemically. Nevertheless, each of these pathogenic states was characterized by an identifiable peripheral blood gene expression distinguishing each state from CTRL (Figure 2-1). Here, we identify the differentially expressed genes driving these signatures.

The gene expression profile characterizing MetS was comprised of many genes involved in innate immune responses (Table 2-2). Gene set 13, Acute Inflammatory Response, featured up-regulation of *CFHR1*, a complement factor gene, and *ORM1*, an

acute phase reactant. Acute phase reactants may be present at increased levels as a consequence of hyperlipidemia-induced liver injury. *CD1D*, involved in antigen presentation of lipids and glycolipids to activate NKT cells, is over-expressed in gene set 316. Under-expression of *TNFAIP3* is found in gene set 407, Negative Regulation of Signal Transduction. This gene encodes a protein that inhibits NF-κB activation and terminates NF-κB responses. Decreased expression of this gene, as with *LILRB4* in RA, limits at least one way in which an immune response is attenuated. Also decreased in expression were two apoptosis-related genes: *DAXX*, involved in TNF-mediated apoptosis, and *MOAP1*, involved in caspase-mediated apoptosis, in the Regulation of Developmental Process gene set. *MAP3K5*, also in gene set 615, shows increased expression. *MAP3K5* activates *MAP2K6* which in turn activates p38 in response to inflammatory cytokines. This pathway was also over-expressed in RA.

Coronary Artery Disease

Peripheral blood gene expression in CAD was also distinguishable from the CTRL cohort (Figure 2-1). This profile is defined by genes that impact activation and expression of NF-κB (Table 2-2). While some genes encoding proteins that impact NF-κB were differentially expressed in RA and MetS, the CAD gene expression profile encompassed a far greater number of NF-κB associated genes. Gene set 499, Positive Regulation of Immune Response, includes the over-expressed genes *IKBKG* and *TLR8*. *IKBKG* is a regulator of the IKK complex, which activates NF-κB; *TLR8* also activates NF-κB as part of the innate immune response. *MAP3K7IP2*, *TNFAIP3* and *TNFRSF10B* are differentially expressed in gene set 636. IL-1 initiated activation of NF-κB is

mediated by *MAP3K7IP2*, *TNFRSF10B* is also an activator while previously mentioned *TNFAIP3,* an inhibitor of NF-κB, is under-expressed in this disease cohort, as well as in MetS. *TRIB3* in gene set 412, Negative Regulation of Transferase Activity, was highly over-expressed. This gene encodes a protein that is induced by NF-κB and acts as a feedback regulator of this transcription factor, thus sensitizing the cells to apoptosis. Downstream effects of activation of NF-κB include increased expression of many genes involved in inflammation and also of genes that protect the immune cells from apoptosis, allowing further expansion of the inflammatory response.

Type 2 Diabetes

The T2D peripheral blood gene expression signature was robust and included many protein-coding genes involved in T cell signaling and function (Table 2-2). Three of these, gene sets 271, 435 and 753, were also significantly differentially expressed in RA and gene sets 435 and 753 specifically showcase the T cell associated genes. Gene set 435, Nucleobase Nucleoside Nucleotide and Nucleic Acid Metabolic Process, includes over-expression of the T cell genes *ILF2,* or nuclear factor of activated T cells, which modulates IL-2 expression, *NFATC4*, a gene involved in the inducible expression of cytokines and *NP*, a gene encoding an enzyme that, when lacking, compromises cell-mediated immunity. In gene set 753, Signal Transduction, the trio of receptors *IL1RL1*, *IL4R* and *IL9R* were over-expressed. *IL1RL1* is a receptor induced by inflammatory cytokines, *IL4R* promotes differentiation of T cells to T helper type 2 cells, and *IL9R* encodes a receptor that supports IL-2 and IL-4 independent growth of the T cell population. This gene set also features decreased expression of the leukocyte

immunoglobulin-like receptors *LILRB2* and *LILRB4*, both of which serve to limit the immune response. *MAPK11,* encoding a protein activated by pro-inflammatory cytokines, is over-expressed in this gene set along with *GPX1*, a glutathione peroxidase. Finally, levels of *TNFRSF13B* transcripts are increased; this gene serves to stimulate lymphocyte function.

A number of other gene sets were significantly differentially expressed in T2D. Cell Cell Signaling, gene set 104, includes up-regulation of the complement component *C1QA* and the chemotaxin *CXCL5*. Gene set 117 features increased expression of *CD276*, another regulator of T cell mediated immunity and *IL2RA*, a gene also over-expressed in RA and involved in proliferation of lymphocytes. The Immune System Process gene set 271 includes many of the previously discussed differentially regulated genes as well as decreased expression of *CTLA4,* a gene encoding a protein expressed on the surface of helper T cells that transduces an inhibitory signal. The gene expression profile of T2D was distinct from CTRL subjects in the differential expression of many genes involved in the activation of and signaling in T cells, reflecting the possibility that components of the adaptive immune system may contribute to the pathogenesis of T2D.

Correlation Among Disease States

To further investigate the overlap in gene expression profiles of the metabolic disorders suggested by hierarchical clustering (Figure 2-2), we explored interrelationships of these profiles in the gene set analysis. To do so, we created a list of gene sets whose average expression level differed significantly from that of CTRL ($p<0.05$) for any of the four comparisons. Next, we assessed the relationships among RA, MetS, CAD and T2D

by estimating pairwise Spearman Correlation coefficients based on the *p*-values for the gene sets derived from the comparison of each state to CTRL. The thickness of the line connecting one state to another is based on the estimated Spearman's correlations (Figure 2-3A). The sole autoimmune disease, RA, showed the lowest correlation with the other diseases. We found the highest degrees of correlation in comparisons among MetS, CAD and T2D demonstrating strong overlap in the peripheral blood gene expression profiles of these inflammatory disease states. Correlation among this trio ranged from Spearman's rho 0.44296 to 0.53772, all with significance of $p<0.0001$. There were 618 genes significantly differentially expressed in 2 or more of MetS, CAD or T2D versus CTRL comparisons (Figure 2-3B).

Within the genes differentially expressed in all three states versus CTRL, *FCGR1A*, an Fc receptor for immunoglobulin-gamma involved in both innate and adaptive immunity, *AGER*, a receptor for the immunogenic advanced glycation end products, the innate immunity-related complement stabilizer *CFP*, and the acute phase reactant, *CP*, were over-expressed. These genes and their related pathways may all lead to activation of the innate immune response. *PPARA*, a peroxisome proliferator receptor, also showed increased expression. *IL2RA*, the NF-κB activator *TNFRSF1A*, and the inflammatory signaling molecule *MAPK11* showed increased expression in both CAD and T2D. IL-1 mediates synthesis of acute phase reactants and the IL-1 receptor associated protein, *IL1RAP*, was differentially expressed in both MetS and T2D along with the NF-κB associated *NFKB2*. Differentially expressed in both CAD and MetS were the innate immune activator *LILRA5*, *MAP3K5* involved in the activation of p38 MAP kinase in response to inflammatory cytokines, and the NF-κB associated *NFKBIB.* Gene

expression profiles of MetS, CAD and T2D were significantly correlated with each other and, to a lesser degree, with RA.

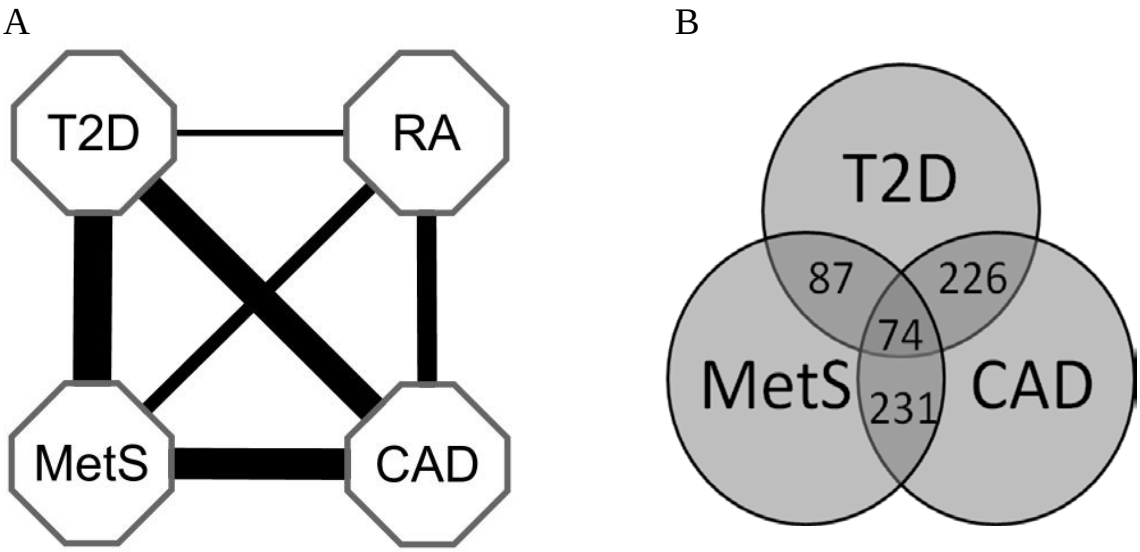

Figure 2-3. Correlative relationships among disease cohort gene expression.
(A) Gene sets that significantly differed in expression versus CTRL were the input for this Spearman's correlation coefficient based diagram. Thickness of the bar represents a combination of Spearman's rho and statistical significance of the correlation. RA= rheumatoid arthritis, T2D= type 2 diabetes, MetS= metabolic syndrome and CAD= coronary artery disease. For the RA-T2D comparison Spearman's rho=0.10396, p=0.0555, RA-CAD rho=0.28462, p<0.0001, RA-MetS rho=0.19942, p=0.0002. T2D compared to CAD rho=0.42389, p<0.0001, T2D-MetS rho=0.53772, p<0.0001 and for the comparison of CAD to MetS rho=0.44296, p<0.0001. (B) A Venn diagram representing the number of genes with significantly different expression in each disease state versus CTRL that overlap among 2 or more of the states.

Given that MetS is a precursor to both CAD and T2D, an analysis was performed to eliminate those gene sets that overlap amongst CAD or T2D and MetS in order to isolate the genes and gene sets that may be involved in progression of MetS to its sequelae. We selected gene sets that were not significantly differentially expressed in MetS but were significantly differentially expressed in CAD or T2D (Table 2-2). As

MetS progresses to its sequelae, CAD, we found differential expression of an increased number of genes involved in activation of and signaling in macrophages. The predominance of genes participating in activation of NF-κB, seen in the comparison of CAD to CTRL, was also found in the comparison of CAD to MetS. The increased expression of monocyte and macrophage related genes can be found primarily in gene set 753, Signal Transduction. *CD14* is a monocyte surface marker, *CXCL14* encodes a chemokine for monocytes, and *MST1R* encodes a protein that serves as the receptor for macrophage stimulating protein. All 3 of these genes were over-expressed in CAD compared to MetS. In addition, three MAP kinases, *MAP2K7*, *MAPK11* and *MAPK13*, were over-expressed in this gene set, all of which are involved in mediating the immune response to pro-inflammatory cytokines. Gene sets 482 and 682 feature a number of genes involved in the activation of the pro-inflammatory transcription factor NF-κB. *CARD14* interacts with *BCL10* to positively influence NF-κB activation; *IKBKG* and *TNFRSF1A* also activate NF-κB. Gene set 372, Negative Regulation of Biological Processes, contains differentially expressed *CLCF1*, a B cell stimulatory cytokine, *F2*, or coagulation factor II, associated with vascular inflammation, and *MPO,* encoding the protein myeloperoxidase, an enzyme found in neutrophils. In addition to the over-expression of NF-κB activating genes, also seen in the direct comparisons of CAD to CTRL and MetS, monocyte and macrophage related genes were also over-expressed in CAD.

Differences in peripheral blood gene expression of T2D as contrasted with MetS were much more subtle than the comparison of CAD with MetS. Of note are *CD276*, *LAT* and *LCK*, in gene sets 271, 478 and 596. *CD276* is involved in regulation of cell-

mediated immune responses in T cells, *LAT* is a component of the cell surface T cell receptor complex, and *LCK* is a protein involved in the maturation and function of T cells. Also significant, *ILF2*, encoding a protein that regulates IL-2 and proliferation of T cells, was over-expressed in T2D relative to MetS. Finally, *CXCL5*, a chemotactic cytokine, was increased in expression in gene set 104, Cell Cell Signaling. T2D and MetS were the two most closely correlated disease states (Figure 2-3A). The gene expression profiles of these two states differ primarily in the over-expression of T cell associated genes in T2D.

PCR Validation

To quantitatively measure differences in transcript levels of a selected group of genes identified by the array analysis, we performed quantitative-reverse transcriptase PCR (RT-PCR). We analyzed 19 of the original 35 samples used for the microarray analysis (group 1). In addition, we obtained 61 independent samples from CTRL, MetS, CAD and T2D subjects (group 2). We determined the fold difference between each experimental group and its own CTRL group, e.g. group 1 or group 2, using the ΔΔCt method (Table 2-3). A 'pooled' p value was calculated by pooling results from groups 1 and 2. From the MetS peripheral blood gene expression profile, *CD1D* also showed increased expression while the decreased expression of *DAXX*, *MOAP1* and *TNFAIP3* was similarly validated. Of interest, our additional analysis demonstrates that *CD1D* and *DAXX* were also differentially expressed in the CAD cohort. Expression of *MOAP1* and *TNFAIP3* was also decreased in all three metabolic cohorts relative to the level in the CTRL cohort in both group1 and group 2. Three genes over-expressed in the CAD

signature were also confirmed by RT-PCR measurements, *CD14*, *CFHR1* and *CXCL14*. These genes also showed significant differential expression in MetS (*CFHR1*), T2D (*CXCL14*) or both (*CD14*). Finally, the differences in expression of *CTLA4*, *GPX1*, *IL4R* and *NP*, from the T2D microarray signature, were confirmed by the RT-PCR experiments. *CTLA4* also displayed decreased transcript levels in CAD and *GPX1* and *IL4R* showed increased and decreased expression, respectively, in all 3 metabolic cohorts. Besides validating results obtained from microarray analyses in independent cohorts by an independent method, these experiments also identify expression patterns of individual genes unique to one or two metabolic disorders or shared by all three metabolic disorders.

Table 2-3. RT-PCR determined ratios[a] of differentially expressed genes

	MetS			T2D			CAD		
Gene	**Grp 1[b]**	**Grp 2[c]**	***p*-value[d]**	**Grp 1**	**Grp 2**	***p*-value**	**Grp 1**	**Grp 2**	***p*-value**
CD14	1.88	1.33	**0.008**	1.68	1.26	**0.005**	1.39	1.39	**0.009**
CD1D	1.59	1.36	**0.006**	1.49	1.35	**0.003**	0.55	0.83	ns
CFHR1	9.21	3.85	**<0.0001**	3.48	3.08	**<0.0001**	1.34	1.48	ns
CTLA4	1.52	0.65	ns	0.21	0.41	**0.002**	0.55	0.58	**0.002**
CXCL14	3.4	0.32	ns	11.82	10.95	**0.002**	10.3	17.95	**0.002**
DAXX	0.33	0.43	**<0.0001**	0.53	0.51	**<0.0001**	0.88	0.72	ns
GPX1	4.75	1.57	**0.002**	2.41	1.61	**0.003**	1.91	2.05	**0.003**
IL4R	0.61	0.51	**<0.0001**	0.26	0.35	**<0.0001**	0.49	0.68	**0.008**
MOAP1	0.42	0.28	**<0.0001**	0.31	0.67	**0.01**	0.72	0.62	**0.03**
NP	2.31	1.06	ns	1.36	1.25	ns	1.82	1.41	**0.02**
TNFAIP3	0.4	0.26	**<0.0001**	0.23	0.27	**<0.0001**	0.97	0.41	**0.03**

[a]ratio= fold change, determined by ΔΔCt calculations, calculated separately for each group versus group-specific CTRLs

[b]Group 1 is composed of 19 samples used in the original gene set analysis (CTRL=4, MetS=6, CAD=3, T2D=6)

[c]Group 2 is an independent set of 61 patient samples (CTRL=16, MetS=16, CAD=13, T2D=16)

[d]*p*-values calculated on groups 1 and 2 pooled data

ns= not significant

Discussion

Our analysis of peripheral blood gene expression in CAD, T2D and their precursor state, MetS, shows that these inflammatory disorders feature unique gene expression signatures. We included individuals with RA in these studies as an example of a disease with a known peripheral blood gene expression profile and for purposes of comparing the metabolic expression signatures to that of an autoimmune disease. As expected, gene expression of the RA cohort was sufficient to distinguish these individuals from CTRL. In each of MetS, CAD and T2D, there were sufficient numbers of genes differentially expressed to cluster the majority of each group away from the CTRL cohort with 100% support. Additionally, when all 4 disease states were included in the analysis, 24/26 subjects from the disease cohorts branched together with 100% support. Within that branch, the RA patients clustered together while the metabolic cohorts showed considerable overlap. Thus, the metabolic cohorts have peripheral blood gene expression signatures that are more similar to RA than CTRL, but also more similar to each other than RA.

The gene expression signature of MetS centers on dysregulation of genes involved in the innate immune response. One component of MetS is hypercholesterolemia, specifically, greater levels of very low density lipoprotein (VLDL). VLDL stimulates release of acute phase proteins from the liver. Activation of the innate immune response in peripheral blood could be a response to increased amounts of circulating VLDL. Fatty acids are known to activate innate immune signaling molecules, like TLR4[151]. The gene expression signature of MetS shares much in common

with that of CAD and T2D; many of the gene sets differentially expressed in the individual comparisons of MetS, CAD and T2D to CTRL also overlap among the three disorders. Spearman's test for correlation showed clear association of the three metabolic disorders, an association that was also significant, but to a lesser extent when correlated to RA. The gene sets and corresponding genes driving this similarity are those associated with activation of the innate immune response, an association not seen in the RA cohort.

In addition to activation of the innate immune response, many genes involved in activation of the pro-inflammatory transcription factor, NF-κB, are differentially expressed in the CAD profile. Comparing CAD and T2D directly to their precursor, MetS, is a more appropriate analysis to determine genes and pathways involved in progression of pre-disease to disease. The comparison of CAD to MetS revealed monocyte and macrophage associated genes are more prominently differentially expressed. In addition to the hyperlipidemia of MetS, diagnosis of CAD indicates the presence of atherosclerotic plaques in the lumen of peripheral blood vessels. CAD gene expression profiles uncovered here reflect systemic inflammation and activation of monocytes. Many of these activated monocytes may migrate from the lumen to become the lipid-filled macrophages seen in the core of these plaques[152]. One possible interpretation of these results is that immunological processes occurring at the site of disease are reflected in peripheral blood.

In the gene expression profile of T2D, a disease that represents more refractory insulin resistance than MetS, we see increased expression of genes associated with activation, signaling and function of T cells. This was also the case in a direct comparison of gene expression between MetS and T2D. Many of these T cell activation genes are

also differentially expressed in RA; however unlike T2D, in RA there is a documented role of T cells in the pathogenesis of disease: as the effector cells of joint-specific destruction[149]. The up-regulation of T cell activation seen in these studies may be a byproduct of enhanced activation of the immune response by adipocytes. Recent studies have shown activated T cells to be present in abundance in visceral adipose tissue of mice with T2D[153].

This independent study also replicates a number of findings in the literature with regards to altered expression of genes in states of insulin resistance and obesity. The monocyte surface antigen CD14, upregulated in MetS, CAD and T2D is also upregulated in mice with insulin resistance[154]. CXCL14 null female mice were protected from obesity-induced hyperglycemia and did not develop insulin resistance[155]. An additional correlation can be found found in a human study in which a SNP in the IL4R gene is associated with increased body mass index[156].

Taken together, our data support a hypothesis whereby MetS produces a state of general systemic inflammation mediated by the innate immune system. This inflammation persists as the pre-disease state progresses to CAD or T2D. Peripheral blood gene expression in CAD and T2D identifies additional immune processes underlying these two disease phenotypes; NF-κB activation in CAD, T cell activation in T2D. Thus, the gene expression profiles of MetS, CAD and T2D present convincing evidence that systemic inflammation is a component of the pathogenesis of all 3 states. Furthermore, this study identifies a minimally invasive system that could be used in a longitudinal study to better understand the progression of MetS to its sequelae.

CHAPTER III

A COMPARISON OF GENOMIC COPY NUMBER CALLS BY PARTEK GENOMICS SUITE, GENOTYPING CONSOLE AND BIRDSUITE ALGORITHMS TO QUANTITATIVE PCR

Abstract

Copy number variants are >1kb genomic amplifications or deletions that can be identified using array platforms. However, arrays produce substantial background noise that contributes to high false discovery rates of variants. We hypothesized that quantitative PCR could finitely determine copy number and assess the validity of calling algorithms. Using data from 29 Affymetrix SNP 6.0 arrays, we called copy numbers using three programs: Partek Genomics Suite, Affymetrix Genotyping Console 2.0 and Birdsuite. We compared array calls at 25 chromosomal regions to those determined by qPCR and found nearly identical calls in regions of copy number 2. Conversely, agreement differed in regions called variant by at least one method. The highest overall agreement in calls, 91%, was between Birdsuite and quantitative PCR. In 38 independent samples, 96% of Birdsuite calls agreed with quantitative PCR. Analysis of three copy number calling programs and quantitative PCR showed Birdsuite to have the greatest agreement with quantitative PCR.

Introduction

Copy Number Variants (CNVs) are defined as amplifications or deletions of >1 kilobase segments of the genome[87,88]. Gene duplications were first identified in the pathogenesis of Charcot-Marie Tooth disease in the 1980s; a copy number (CN) amplification of the PMP22 gene was shown to be sufficient to cause disease[112]. These regions of variance were thought to be rare and when the human genome was published, variance amongst humans was primarily attributed to base-pair level single nucleotide polymorphisms (SNPs)[75,157]. However, CNVs were discovered to be present and widespread in the genome shortly thereafter[87,88]. These variants are generated during normal recombination events, leading to inherited CNVs, as well as somatically throughout life in rapidly dividing cells[96,158,159]. CNVs can directly influence gene expression through dosage effects where more copies of the gene produce greater expression, and also by altering transcriptional regulation in the genome, both in the region of variance itself and also in regions up to 1 megabase away[98,160,161].

CNVs can be detected by fluorescence in situ hybridization, bacterial artificial chromosome arrays, genome-wide SNP arrays or direct quantitative PCR (qPCR) in a genomic region of interest. One example of a genome-wide array is the Affymetrix SNP 6.0 array, with close to 1 million probes for determining SNPs across the genome and an additional ~ 1 million probes specifically designed to assess CN. Data from these arrays can be transformed into CN using any of a number of methods, including defined threshold intensity cut-offs and complex statistical algorithms like circular binary segmentation and the Hidden Markov Model[102]. These calling methods are built in to user

accessible programs like Partek Genomics Suite, Affymetrix's Genotyping Console 2.0 and Copy Number Analysis Tool, and Birdsuite software developed by the Broad Institute at Harvard, among others[162]. A recent analysis evaluating the performance of seven CN calling algorithms- circular binary segmentation[163], CNVFinder[164], cnvPartition, gain and loss of DNA[165], Nexus segmentation methods Rank and SNPRank, PennCNV[166] and QuantiSNP[167]- found QuantiSNP outperformed other methods and had the highest statistical power to detect CNVs[168]. However, this comparative analysis was based on consensus of calls amongst the methods and did not assess a non-array reference, like qPCR, that might determine the accuracy of the calls.

Due to concerns about accuracy when only one calling method is used, CNVs have been associated with a number of diseases and states based on the use of multiple algorithms with a consensus call made for each genome region[169], or from just one algorithm paired with additional validation like qPCR or multiple ligation-dependent probe amplification methods[73,170,171,172,173]. In addition to the programs and methods already mentioned, new methods continue to be introduced in the literature[174,175].

We hypothesized that qPCR could finitely determine CN and through this process, assess the validity of a calling algorithm. To test this hypothesis, we took data from 29 Affymetrix SNP 6.0 arrays and called CNs across the genome using three separate programs: Partek Genomics Suite, Affymetrix Genotyping Console 2.0 and Birdsuite. We compared the array calls at 25 individual chromosomal regions with CN calling in the same genomic DNA samples by qPCR.

Materials and Methods

Patient Recruitment

Patients were recruited by the Clinical Research Center at Vanderbilt University. These studies were approved by the Institutional Review Board of Vanderbilt University and all subjects provided written informed consent.

Affymetrix SNP 6.0 Arrays

Peripheral blood was drawn into a Vacutainer Venous Blood Collection Tube (BD Catalog #367861) containing EDTA. Equal volume of lysis buffer (0.32M Sucrose, 10mM Tris-HCL, 5mM MgCl2, 0.75% Triton X-100, pH 7.6) and 2X volumes of dH20 were added to each. Samples were centrifuged and resuspended in lysis buffer. After a second centrifugation, the pellet was resuspended in proteinase K buffer (20mM Tris-HCl, 4mM Na_2-EDTA, 100mM NaCl, pH 7.4) and proteinase K (20mg/ml) was added to the solution. Samples were incubated for 1h at 55°C, cooled on ice and 5.3M NaCl was added. Samples were centrifuged, supernatants kept and added to cold isopropanol and incubated for 30 minutes. Finally, genomic DNA was centrifuged and the pellet was washed twice with 70% ethanol. Genomic DNA was dissolved in Tris-HCl (pH 8.0) and hybridized to the Affymetrix Genome-Wide Human SNP Array 6.0 (Santa Clara, CA) according to the manufacturer's protocol. Following scanning, arrays were checked for quality using Affymetrix Genotyping Console. Arrays with a Contrast QC less than 0.4 were removed from further analysis.

Copy Number Analysis

Genotypes and CN were called using three different methods. The data were loaded into Partek Genomics Suite, quantile normalized and compared to the HapMap 6.0 baseline. CNV regions were called based on the presence of at least 3 consecutive probe sets. Data were also loaded into the Affymetrix program Genotyping Console 2.0. CN was determined with reference to the GenomeWideSNP_6.hapmap270 file and CNVs were similarly called based on variance of at least 3 consecutive probes. Finally, data were inputted into Birdsuite v.1.5.3 and variant regions were called without a pooled reference file. As a further quality control step for Birdsuite, arrays with an overall call rate less than 98% were discarded from further analysis.

Quantitative PCR Experiments

To validate the CN of variant regions from the Affymetrix chip, primer assays were ordered from Applied Biosystems, either custom designed or selected from their inventoried stock of assays, all designed specifically to detect genomic CN (Supplementary Table 3-2). Reactions were run with 20ng genomic DNA per the standard Applied Biosystems protocol in a 7300 Real Time PCR System. All samples were run in triplicate with a multiplexed RNase P or Hemoglobin-beta reference assay and CN was called using ΔΔCt values calculated in Applied Biosystem's CopyCaller v.1.0. In cases where a calibrator sample with CN of 2 was not known, plates were calibrated to an average CN=2.

Results

Genomic DNA samples from 77 individuals were hybridized on Affymetrix SNP 6.0 Arrays. 29 of these samples were analyzed for CN using the Partek Genomics Suite, Genotyping Console 2.0 (GTC) and Birdsuite (Supplementary Table 3-1). Of note, both Partek and GTC CN calls were determined using a pooled HapMap comparison file. qPCR analysis was performed to determine CN at a total of 25 individual genomic regions across 12 chromosomes. The results were compared to the 3 sets of genome-wide calls made from the arrays (Supplementary Table 3-2).

A number of regions were identified by one or more of the algorithms to have CN of 2 in all samples tested. We probed 16 of these "invariant" regions by qPCR and compared the results of the calls in each sample by each method (Table 3-1). There was vast agreement in CN calls in these regions. Of note, qPCR called 4 samples variant at chromosome 2 that were called CN of 2 by all three algorithms. Additionally, in a region on chromosome 8, Partek and GTC called all 8 samples a CN of 2 while Birdsuite called 2 samples variant. Those same 2 samples were also found to be variant by qPCR. Additionally, one sample was called variant by GTC on chromosome 9 but invariant, or CN of 2, by all other methods. All together, seven sample-region pairings were called variant by just one or two methods and invariant by the others while 209 sample-region pairings were uniformly called CN of 2 by Partek, GTC, Birdsuite and qPCR, representing nearly 97% agreement amongst all methods of CN calling in these 16 regions.

Table 3-1. Copy number calls at invariant regions of the genome

Region	Partek	GTC[a]	Birdsuite	qPCR
2:240,032,091	28/0[b]	28/0	28/0	24/4
3:180,366,781	27/0	27/0	27/0	27/0
6:326,150	5/0	5/0	5/0	5/0
7:11,288,419	17/0	17/0	17/0	17/0
7:24,002,710	24/0	24/0	24/0	24/0
8:51,195,001	8/0	8/0	6/2	6/2
9:16,930,899	25/0	24/1	25/0	25/0
13:53,784,055	9/0	9/0	9/0	9/0
13:56,713,547	9/0	9/0	9/0	9/0
15:32,555,299	8/0	8/0	8/0	8/0
16:3,104,307	9/0	9/0	9/0	9/0
16:4,280,826	8/0	8/0	8/0	8/0
16:18,557,305	9/0	9/0	9/0	9/0
20:28,068,523	6/0	6/0	6/0	6/0
20:29,271,114	8/0	8/0	8/0	8/0
22:47,400,722	16/0	16/0	16/0	16/0

[a]GTC=Genotyping Console 2.0

[b]Values are represented as "#samples with CN=2" / "#samples with CN=non 2"

Additional regions were identified as variant, containing numerous CNVs amongst the 29 samples. Nine of these regions were investigated by qPCR, 184 sample-region pairs in total, and the results produced by the three CN calling algorithms were compared by CN class, 0, 1, 2, 3 and 4 (Table 3-2). Region 1 on chromosome 2 produced an identical group of CN calls amongst each of the 4 methods. GTC, Birdsuite and qPCR also produced identical CN calls in Region 2. Additional regions 3-9 did not show such similarity in CN calls between the 4 methods but these comparisons suggested that agreement was highest when Partek calls were compared to GTC calls or when Birdsuite calls were compared to qPCR. Regions 3, 4 and 8 produced a very similar breakdown of calls in Partek and GTC while region 7 was nearly identical, with 7 samples being called amplifications by both programs. Regions 3, 4, 6, 7 and 9 showed similar calls by both

Birdsuite and qPCR. These analyses indicate that some patterns of agreement were observed amongst the different methods of CN calling.

Table 3-2. Comparison of copy number calls at variant regions

Region 1
Chr2:242,648,367

	0	1	2	3	4
Partek	0[b]	4	24	0	0
GTC[a]	0	4	24	0	0
Birdsuite	0	4	24	0	0
qPCR	0	4	24	0	0

Region 2
Chr3:53,010,599

	0	1	2	3	4
Partek	0	1	24	0	0
GTC	1	3	21	0	0
Birdsuite	1	3	21	0	0
qPCR	1	3	21	0	0

Region 3
Chr7:133,441,893

	0	1	2	3	4
Partek	0	1	5	0	0
GTC	0	1	4	1	0
Birdsuite	1	3	2	0	0
qPCR	1	3	2	0	0

Region 4
Chr8:144,776,429

	0	1	2	3	4
Partek	0	0	8	0	0
GTC	1	0	7	0	0
Birdsuite	1	2	5	0	0
qPCR	1	2	5	0	0

Region 5
Chr19:56,824,268

	0	1	2	3	4
Partek	0	0	13	13	0
GTC	0	0	17	3	6
Birdsuite	0	0	26	0	0
qPCR	1	6	19	0	0

Region 6
Chr19:56,838,253

	0	1	2	3	4
Partek	0	0	7	6	0
GTC	0	0	6	3	4
Birdsuite	0	2	11	0	0
qPCR	0	2	11	0	0

Region 7
Chr20:1,522,539

	0	1	2	3	4
Partek	0	4	13	7	0
GTC	0	4	13	0	7
Birdsuite	17	7	0	0	0
qPCR	18	6	0	0	0

Region 8
Chr22:22,643,636

	0	1	2	3	4
Partek	0	0	23	5	0
GTC	1	0	24	3	0
Birdsuite	12	11	5	0	0
qPCR	0	0	28	0	0

Region 9
Chr22:22,713,888

	0	1	2	3	4
Partek	0	0	14	12	0
GTC	1	1	16	6	2
Birdsuite	2	13	11	0	0
qPCR	2	14	10	0	0

[a]GTC=Genotyping Console 2.0
[b]Values are numbers of samples called each CN at each region.

To determine exact agreement among the CN calling methods, 400 CN calls were compared on a sample-by-sample basis to determine agreement of each CN state (Figure 3-1). As previous analyses indicated, the highest agreement in every comparison was seen at CN=2 among the individual CN states (0, 1, 2, 3, and 4). These agreements ranged from 82% to 96% and greatly influenced the overall agreements of each comparison. Discordance among calls from each method was found by comparing the variant calls (CN of 0, 1, 3 or 4).

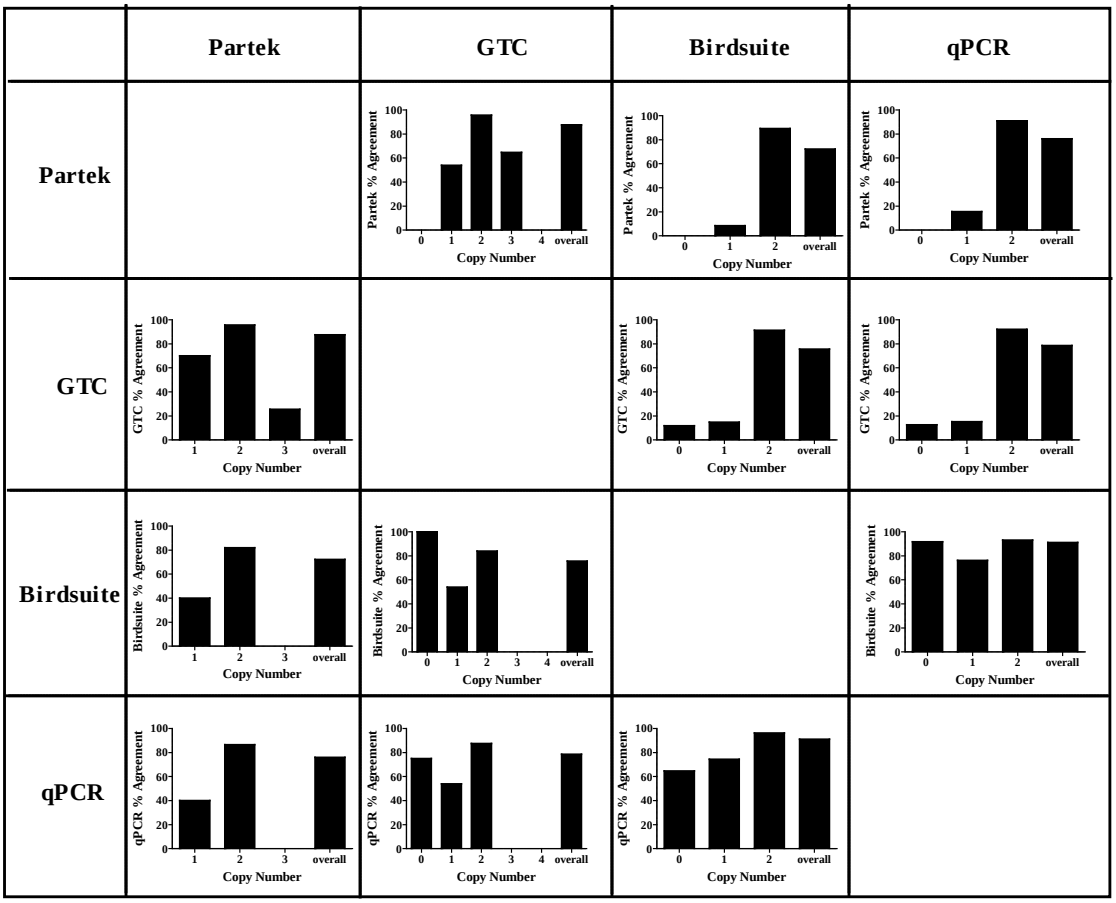

Figure 3-1. Agreement of CN calls made by Partek, GTC, Birdsuite and qPCR.
CN calls were analyzed on a sample-by-sample basis across 25 individual chromosomal regions. Results were sorted according to the CN called by the method named at the top of each column. Descending in each column are method-by-method comparisons. A total of 400 CN calls were considered in this analysis. Results are expressed as % agreement between any two methods of CN call determinations.

When Partek called a CN of 1, GTC also called the sample a CN=1 70% of the time. However, when GTC called a sample region CN=1, Partek correctly called that region a CN=1 54% of the time. There was no agreement between GTC and Partek at CN of 0 because Partek did not call any CN=0 in any of the tested regions. Partek showed less than 50% agreement with variant calls in both Birdsuite and qPCR. The overall agreement of GTC with Partek was 88%, between Birdsuite and Partek was 72% and qPCR CN calls agreed with Partek CN calls 76% of the time.

When GTC called a CN of 0, Birdsuite also called a 0 100% of the time while the agreement with qPCR was 75%. At CN of 1, Birdsuite and qPCR had an identical call in less than 60% of samples. Conversely, when Birdsuite or qPCR made a CN call of 0 or 1, GTC reported the same call in those samples less than 20% of the time. Overall agreement of Birdsuite with GTC was 76% and qPCR and GTC agreed in 79% of the samples.

When Birdsuite called a CN of 0 or 1, the agreement with Partek was 0% and 9%, respectively. Birdsuite variant calls agreed with GTC's calls at slightly higher rates, 12% for CN=0 and 15% for CN=1. The agreement between Birdsuite and qPCR, however, was 65% for CN of 0 and 75% for CN of 1. Of note, the majority of the disparate calls in this comparison came from region 8 (Table 3-2), where Birdsuite called 23 samples to be CN deletions while qPCR determined them to be CN=2. While GTC and qPCR showed high agreement at CN of 0, the agreement at CN=1 was 54%. The Birdsuite agreement with qPCR in CNV sample regions were thus the highest seen among any comparison of array-based calls with qPCR. The overall agreement of qPCR with Birdsuite was also the highest, at 91%.

Finally, when CN calling from array-based algorithms were compared to qPCR, GTC and Partek both showed variant agreements less than 20% of the time, while Birdsuite agreed with 92% of qPCR calls at CN=0 and 76% of qPCR calls at CN=1. Similar to the inverse comparison of qPCR calls to Birdsuite, when Birdsuite variant calls were compared to qPCR, region 5 (Table 3-2) showed 6 samples that were determined to be CN deletions by qPCR and called CN of 2 by Birdsuite, accounting for a large portion

of the 24% error in calls at CN=1. Overall, the highest agreement was found between qPCR and Birdsuite.

We next assessed if the high percent agreement between Birdsuite and qPCR was reproducible in a second independent group of samples. Data from 38 additional Affymetrix SNP 6.0 arrays were analyzed by Birdsuite to determine CN calls across the genome (Supplemental Table 3-1). qPCR reactions were performed using 18 different assays investigating regions on 10 chromosomes to determine CN at each region. A total of 387 comparisons were made in this step (Table 3-3 and Figure 3-2). A total of 14 Birdsuite calls in 7 genomic regions did not agree with the CN call made by qPCR (Table 3-3). Six of these disparate calls were CN=2, 7 of CN=1 and 1 of CN=3. Overall agreement at each CN was also determined (Figure 3-2). Birdsuite and qPCR agreed on 100% of CN=0, 87% of CN=1 and 98% of CN=2. Of note, there was 0% agreement in 1 sample called a CN=3 by qPCR. The overall agreement of Birdsuite with qPCR was 96%, better than the overall agreement rate from the first analysis (91%).

Table 3-3. Birdsuite agreement with qPCR calls in 18 genomic regions

Region	0	1	2	3
2:242,648,367	0 (0)	1 (0)	36 (0)	0 (0)
3:53,010,599	1 (0)	2 (1)	30 (0)	0 (0)
3:180,366,781	0 (0)	0 (0)	16 (0)	0 (0)
7:11,288,419	0 (0)	0 (0)	26 (0)	0 (0)
7:24,002,710	0 (0)	0 (0)	16 (0)	0 (0)
7:133,441,893	2 (0)	14 (0)	6 (0)	0 (0)
8:51,195,001	1 (0)	11 (2)	13 (0)	0 (0)
8:144,776,429	1 (0)	5 (2)	10 (2)	0 (0)
9:16,930,899	0 (0)	0 (0)	16 (0)	0 (0)
13:53,784,055	0 (0)	0 (0)	18 (0)	0 (0)
13:56,713,547	0 (0)	0 (0)	13 (4)	0 (0)
13:71,376,533	7 (0)	12 (1)	1 (0)	0 (0)
15:32,555,299	0 (0)	3 (0)	16 (0)	0 (0)
16:3,104,307	0 (0)	0 (0)	18 (0)	0 (0)
16:4,280,826	0 (0)	0 (1)	15 (0)	0 (0)
16:18,557,305	0 (0)	0 (0)	26 (0)	0 (0)
20:29,271,114	0 (0)	0 (0)	8 (0)	0 (1)
22:47,400,722	0 (0)	0 (0)	29 (0)	0 (0)

[a]Values are represented as the number of calls that "agree(disagree)"

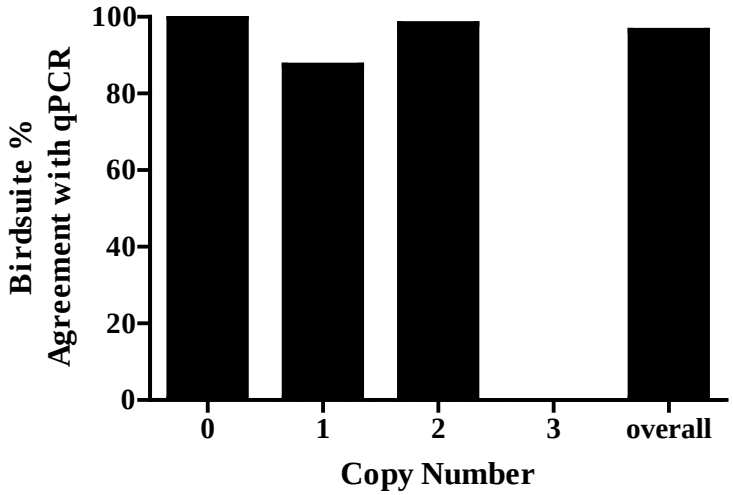

Figure 3-2. Agreement between CN calls made by Birdsuite and qPCR.
Birdsuite CN calls were compared to qPCR CN calls at 18 distinct chromosomal regions on 10 chromosomes. A total of 387 data points were considered in the analysis. At CN=0, 12/12 samples agreed, at CN=1, 48/55 samples agreed, at CN=2, 313/319 samples agreed and at CN=3, 0/1 sample agreed. Results are expressed as % agreement between any the two methods of CN call determinations.

Discussion

A total of 77 peripheral blood genomic DNA samples were analyzed for CN on Affymetrix SNP 6.0 Arrays. CN calls for 29 of these samples were determined by three different methods: Partek Genomics Suite, GTC and Birdsuite. Calls at 25 genomic regions were also determined by qPCR in these same samples. Comparison of these CN calls shows that all 4 methods agreed when the CN call is 2. However, there is considerably less agreement when the CN calls identify variant regions or CNVs. One way to determine one singular CN call for each region would be pooling the array CN calls from each algorithm to arrive at a consensus call. However, the disagreement

amongst variant calls by Partek, GTC and Birdsuite seen in this sampling prohibit arriving at a clear consensus. When each of the 4 methods was compared to the others, the highest agreement both in variant calls and overall calls was between Birdsuite and qPCR.

CN calls made by each algorithm- Partek, GTC and Birdsuite, on each array are subject to a number of assumptions. CN calls are calculated in GTC and Partek by using a reference file or baseline. This reference file is generally composed of pooled CN data such that the average intensity of the group is assumed to be CN of 2. However, in the 77 arrays analyzed we discovered numerous regions to be variant in greater than 80% of samples. Pooling these arrays and assuming a CN of 2 would therefore skew results. In contrast, Birdsuite uses a unique method to determine CN. The Broad Institute has previously characterized copy number polymorphisms by determining those CNVs present in greater than 5% of the HapMap population[176]. Birdsuite uses known intensity value-CN references at the 1,320 copy number polymorphisms to infer CN in the remaining portions of the genome[162]. However, no algorithm can completely escape the problem of background intensity on the array and the risk for type I and type II errors that come with the sampling of intensity values at nearly 2 million probes.

CN determined by qPCR is not without assumptions. Calls are made using the ΔΔCt calculation with the first comparison coming between the test assay Ct value and a multiplexed reference assay Ct value. Reference assays exist for genes known to be CN invariant, or to always have exactly 2 copies of the gene in the genome. The second comparison is made between the test assay-reference assay value and that same value for a calibrator sample, known to have a CN=2 in the test region. If the calibrator sample is

not a CN of 2, the data would be skewed in the direction of the actual CN of the calibrator sample. qPCR, however, does not have the problem of additional background noise and is also immune to multiple sampling errors. For these reasons, qPCR is considered to be the standard in determining CN.

The algorithm employed by Birdsuite to call CNs across the genome closely agrees with the qPCR determinations of CN. When all 787 comparisons from these data are considered, the overall agreement is 94%. For this reason, the use of the Birdsuite algorithms, in combination with PCR validation, generated the most reproducible CN calls in this group of patient samples. Of note, more recent versions of Genotyping Console now employ the Birdsuite algorithms to determine CN.

CHAPTER IV

GENOME-WIDE ANALYSIS OF COPY NUMBER VARIATION IN TYPE 1 DIABETES[a]

Abstract

Type 1 diabetes tends to cluster in families, suggesting there may be a genetic component predisposing to disease. However, a recent large-scale genome-wide association study concluded that identified genetic factors, single nucleotide polymorphisms, do not account for overall familiality. Another class of genetic variation is the amplification or deletion of >1 kilobase segments of the genome, also termed copy number variations (CNVs). We performed genome-wide CNV analysis on a cohort of 20 unrelated adults with type 1 diabetes and a control cohort of 20 subjects using the Affymetrix SNP Array 6.0 in combination with The Birdsuite copy number calling software. We identified 39 CNVs as enriched or depleted in type 1 diabetes versus control. Additionally, we performed CNV analysis in a group of 10 monozygotic twin pairs discordant for type 1 diabetes. Eleven of these 39 CNVs were also respectively enriched or depleted in the Twin cohort, suggesting that these variants may be involved in the development of islet autoimmunity, as the presently unaffected twin is at high risk for developing islet autoimmunity and type 1 diabetes in their lifetime. These CNVs include a deletion on chromosome 6p21, near an HLA-DQ allele. CNVs were found that were both enriched or depleted in patients with or at high risk for developing type 1

[a] This chapter has been published in: PLoSOne. 2010 Nov 15; 5(11):e15393

diabetes. These regions may represent genetic variants contributing to development of islet autoimmunity in type 1 diabetes.

Introduction

Type 1 diabetes (T1D) results from immune-mediated selective destruction of pancreatic islet cells resulting in insulin deficiency and hyperglycemia[6,7]. Symptoms of polydipsia, polyuria, polyphagia and weight loss manifest when significant numbers of islet cells have been destroyed. However, antibodies to islet autoantigens can be detected in peripheral blood prior to clinical disease[6,21]. With early diagnosis of disease or assessment of risk, immune therapy may impede islet destruction and preserve insulin production, delaying onset of clinical manifestations[7].

Another component of T1D that aids in our understanding of the disease and assessment of risk is genetic inheritance. A long-term (up to 40 year) study of twin pairs in Finland revealed a monozygotic (MZ) pairwise concordance for T1D of 27.3% while the concordance for dizygotic (DZ) twins was 3.8%[20]. The impact of genetics was further made clear in this study because upon diagnosis of T1D in one twin, the length of time to diagnosis in the other twin in the concordant pairs was a maximum of 6.9 years in MZ twins and 23.6 years in DZ twins[20]. In addition to measuring incidence of T1D in twin studies, islet antigen-specific autoimmunity can also be determined. As a precursor to T1D, autoimmunity is defined as the presence of antibodies to islet autoantigens in sera[177]. In another study, 83 unaffected monozygotic twins were followed for nearly 44 years and incidence of autoimmunity or diagnosis of T1D was recorded. This study

showed a 65% cumulative incidence of T1D by 60 years of age and more than 75% tested positive for an islet autoantibody during the course of the study. Once autoimmunity was established, the risk of diabetes was 89% within 16 years of the first positive autoantibody test.

Clearly genetics play an important role in the T1D disease process as both MZ and DZ twins have the same environmental exposures but different concordance rates and length to diagnosis of the second twin. Numerous genes have been associated with T1D, the most significant being the HLA region on chromosome 6[178]. More than 90% of type 1 diabetics carry HLA alleles DR3-DQ2 or DR4-DQ8 compared to no more than 40% of the general population[179]. Alleles at HLA-DQB1 are known to be, in part, protective[180]. Single nucleotide polymorphisms (SNPs) are also associated with T1D. A recent genome-wide association study of approximately 2,000 patients with each of 7 common, chronic diseases, including T1D, and 7,000 shared controls confirmed the association of SNPs in 5 previously identified regions with T1D and discovered 5 novel associations. However, the authors concluded that these regions, with the exception of the HLA on chromosome 6, confer only modest effects on T1D, and "the association signals so far identified account for only a small proportion of overall familiality"[26]. These results suggest that additional genetic variants contribute to inheritance of T1D.

Another class of genetic variation is the amplification or deletion of >1 kilobase segments of the genome, also called copy number variations (CNVs)[87,88]. Gene duplications were first identified in the pathogenesis of Charcot-Marie Tooth disease in the 1980s; a CN amplification of the PMP22 gene was shown to be sufficient to cause disease[112]. These regions of variance were thought to be rare and when the human

genome was published, variance amongst humans was primarily attributed to base-pair level SNPs[75,157]. However, CNVs were discovered to be present and widespread in the genome shortly thereafter[87,88]. These variants are generated during normal recombination events, leading to inherited CNVs, as well as somatically throughout life in rapidly dividing cells[96,158,159]. CNVs can directly influence gene expression through dosage effects where more copies of the gene produces greater expression, and also by altering the transcriptional regulation of the genome, both of the region of variance itself and regions up to 1 megabase away[98,160,161].

Monozygotic twins discordant for disease represent a controlled population in which to identify potentially disease-associated CNVs. Monozygotic twins do not have identical genetic sequences and are known to vary in CNVs and at the epigenetic level[170,181,182,183]. Differences may arise during prenatal cell division or post-natally in continuously dividing cells like lymphocytes. The latter would result in CNVs that not only differ from the co-twin but also from CNVs in other cells and tissues of the body. In the case of disease discordant monozygotic twins, if a CNV were associated with a certain disease, we presume the twin affected by the disease would have the variant and the unaffected twin would not. A study of nine MZ twin pairs discordant for Parkinson's disease identified 35 regions of variance present in only the affected twin of at least four of those pairs, confirming the hypothesis that MZ twins differ in CNVs and that these regions may be involved in the development of disease, as evidenced by the presence of specific CNVs in multiple affected twins[181].

There are numerous other diseases and states associated with differences in CNVs, among them schizophrenia and adult deficit hyperactivity disorder[181,184,185]. But

not all CNV associations are with neurologic or behavioral diseases. Recent studies have shown additional functional implications of CNV and disease, notably in studies of CNV of the Fc-gamma receptor and the autoimmune disease systemic lupus erythematosus (SLE). Patients with SLE are more likely to have fewer copies of *FCGR3B*, encoding a protein involved in the uptake and clearance of immune complexes[186].

We hypothesize that CNVs contribute to susceptibility to and/or protection from T1D. To test this hypothesis, we performed genome-wide analysis on a cohort of 20 unrelated adults diagnosed with T1D and 20 unrelated control (CTRL) subjects to identify CNVs either enriched or depleted in the T1D cohort compared to CTRL. We then looked at the frequency of these variants in a second cohort of 10 MZ twin pairs disease discordant for T1D. The frequencies of the CNVs of interest did not differ from the affected twin subset to the unaffected twin subset. However, because of the high lifetime incidences of autoimmunity and/or T1D in the unaffected twins, the 10 twin pairs were considered as a single cohort with or at high risk for T1D[21]. This analysis identified 5 CNVs enriched and 4 CNVs depleted in both the T1D and Twin cohorts.

Materials and Methods

Ethics Statement

These studies were approved by the Institutional Review Board of Vanderbilt University and all subjects provided written informed consent. Monozygotic twin blood samples and family history information were provided with written informed consent

using protocols and consent forms approved by the Colorado Multiple Institutional Review Board.

Patient Recruitment

Diabetes is defined by the WHO criteria of classic symptoms of diabetes (polydipsia, polyuria, polyphagia and weight loss) and a plasma glucose >200 mg/dl, a fasting plasma glucose of >126 mg/dl or a 2 h plasma glucose during an oral glucose tolerance test of >200 mg/dl[6]. T1D is differentiated from type 2 diabetes by a number of criteria- history, clinical presentation and laboratory findings, including antibody testing when available. *Control* patients have never received a diagnosis of a chronic disease or syndrome and are not currently taking medication for any illness or condition. *Rheumatoid arthritis* is defined by the American College of Rheumatology Criteria. Patients displayed four or more of the following symptoms for greater than 6 months: morning stiffness, swelling in 3 or more joints, swelling of finger and/or wrist joints, symmetric swelling, rheumatoid nodules, positive rheumatoid factor, or radiographic erosions in the hand and/or wrist[139]. *Multiple Sclerosis* patients were recruited with the following characteristics: diagnosis of relapsing remitting MS (RRMS) based upon the revised McDonald criteria[187,188], no prior cytotoxic treatments that might induce DNA damage, no family history of MS in either first or second degree relatives, and age between 25-35 (to restrict the possibility of age-related somatic mutations).

Ten pairs of monozygotic twins were selected from the Barbara Davis Center Twin Family Study, an ongoing, long-term follow up study of initially unaffected twins of patients with type 1 diabetes. Twins are ascertained through various sources,

including the Barbara Davis Center Clinic, the Joslin Diabetes Clinic, the Diabetes Prevention Trial, TrialNet, and other physician and self-referrals. Family history of diabetes and other autoimmune diseases is collected at enrollment and updated over time. Serum and DNA samples are collected from twins and other family members. Serum is tested for the presence of islet autoantibodies as well as celiac and adrenal autoantibodies. Autoantibody testing is repeated for unaffected twins for as long as they remain in the study or until they develop diabetes. Twin zygosity is confirmed by testing a panel of 16 microsatellite markers. Twin DNA samples included in the present study were collected within 14 months of diagnosis of the affected twin, and at approximately the same time (within 1 week) for the two twins of each pair.

Genomic DNA samples from 73 patients with T1D, comprising the independent cohort for qPCR analysis, were obtained from Coriell Cell Repository, repository number 65895.

Affymetrix Copy Number Variation Experiments

Please refer to methods in Chapter III (page 63). Genotypes and CNs were called using Birdsuite v1.5.3. As a further quality control step, arrays with an overall call rate less than 98% were discarded from further analysis.

Copy Number Analysis

Genomic regions with a Birdsuite CN call confidence value less than 5 were merged to the adjacent region with a confidence score greater than 5, assuming the CN of that confident region. Next, each genome was narrowed down to a list of genomic

variants with confidence scores greater than 5. Regions of CN=2 were discarded. These lists were merged with the list of CNPs and regions of variance not represented by a CNP were denoted as "novel"[176]. Next, CNVs that were present in greater than 40% of the T1D group at a 1.5 fold change greater frequency as compared to the CTRL group were determined to be enriched and CNVs present in greater than 40% of the CTRL group at a 1.5 fold change greater frequency as compared to the T1D group were determined to be depleted. For validation of these CNVs whose presence or absence may be associated with diabetes, an identical analysis was performed on the discordant twin cohort, as compared to CTRL and with the additional step of comparing the affected twins versus their unaffected co-twin pair to determine any variants present differentially within the group. CNVs enriched or depleted in both the twins and unrelated T1D adults were selected for further analysis. Chi-square analyses were performed on each CNV of interest in each cohort versus CTRL based on the presence or absence of variance.

We assessed statistical significance for the observed overlapping CNVs in both T1D cohorts relative to the CTRL group using a permutation test with 1000 permutations. Briefly, keeping the number of patients fixed in each of the three groups, we randomly permutated group status for the samples (so that they were re-assigned to different groups) and re-calculated the number of enriched CNVs in both disease groups relative to CTRL. This process was repeated 1000 times. The p-value for the number of observed overlapping CNVs (i.e. 10) was estimated by the number of permutations with 10 or more overlapping CNVs divided by the total number of permutations.

Quantitative PCR Experiments

See methods in Chapter III (page 64).

Results

We sought to determine if CNVs are associated with T1D by performing genome-wide CNV analysis on a cohort of 20 patients with T1D and 20 CTRL patients using the Affymetrix SNP Array 6.0. An additional cohort of 10 monozygotic twin pairs discordant for T1D was analyzed for validation purposes. Quality of the hybridization, as defined by Affymetrix in the Genotyping Console as a contrast QC <0.4, was assessed and 1 CTRL sample failed prior to CN analysis.

Of primary importance in the analysis of these data was the validity of our CN calling algorithm. Raw data from all 3 cohorts, 59 Affymetrix arrays in total, were inputted into the Birdsuite programs and CNs were called across the genome. The Birdsuite software determined integer CNs of predefined regions of common variance (copy number polymorphisms, CNPs) and employed a more complex, multi-dimensional model to identify rare variants[162]. Output files contain CN values across the chromosome with a confidence score of each individual call (Supplemental Table 4-1). Genome wide call rates were also estimated for each individual sample. Two samples from the unrelated adult T1D cohort failed a quality control checkpoint of call rate greater than 98%. The remaining 57 arrays had call rates $\geq$ 98.6%. In a similar analysis to those described in Chapter III, we compared CNs determined by the Birdsuite analysis to CNs determined by quantitative PCR (qPCR) in 37 samples of genomic DNA across 5 distinct

chromosomal regions (Supplemental Table 4-2). For qPCR experiments, Applied Biosystem's CopyCaller1.0 program determined a non-integer CN based upon the ΔΔCt calculation and then predicted an integer CN, each with an associated confidence value. For 185 separate experimental points, there was > 96% agreement in CN determinations made by the Birdsuite analysis and the qPCR analysis (Figure 4-1).

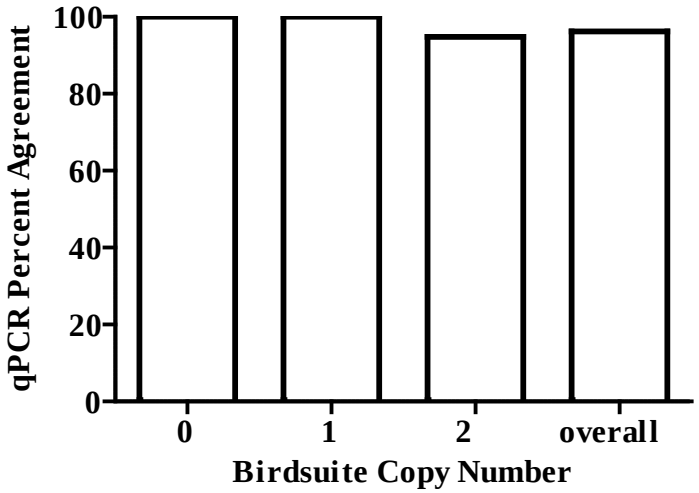

Figure 4-1. Percent agreement between Birdsuite copy number calls and qPCR. Percent agreement between the Affymetrix array CNs, as determined by the Birdsuite software, and qPCR CN, determined using ΔΔCt calculations, is plotted for each CN class. Percents are based on 214 comparisons from CNs for 37 samples on 6 distinct chromosomal regions (7q33, 8q11, 8q24, 15p13, 16p12). For CN 0, 6/6 samples agreed (100%). For CN 1, 45/45 samples agreed (100%). For CN 2, 127/134 of samples agreed (94.7%). The overall agreement is 178/185 samples (96.2%).

For the analysis, we first catalogued all confident, variant CN calls on chromosomes 1-22 within the framework of known copy number polymorphisms (CNPs)[176]. CNPs are regions of CN variance present in greater than 1% of the 270 HapMap samples, resulting in a library of 1,320 CNPs. Novel CNVs not represented in the CNP library were also identified and included in the analysis.

A single CNV is capable of causing disease. In the case of Charcot Marie Tooth disease type 1, 70% of patients have one singular pathogenic variant[112,189]. In a more common disease, like T1D, we hypothesized that a single undiscovered variant would not be present and pathogenic at a percentage as high as 70%, rather we set the threshold of variance within each diabetes group at roughly half that, or 40%. Additionally, to ensure selection of variants differentially expressed between the two groups, we further limited the classification of enrichment in T1D to those CNVs present at a 1.5 fold greater frequency than CTRL. Conversely, a CNV was classified as depleted in T1D if it was present in >40% of the CTRL cohort at a 1.5 fold greater frequency than T1D.

Variants in the T1D group were compared to those in the CTRL group and 18 CNPs present in > 40% of the T1D cohort at a 1.5 fold greater frequency than the CTRL cohort were identified as enriched in T1D. Conversely, 20 CNPs and 1 novel CNV were depleted in the T1D cohort, defined as a variant present in >40% of the CTRL cohort at a 1.5 fold greater frequency than the T1D cohort. These 39 CNVs were then studied in a second cohort.

The Affymetrix chip determines CN based on values of nearly 1,000,000 probes in the genome, resulting in a high probability for both type I and II errors. To help control for these errors, we performed genome-wide CN analysis in a second cohort of patients, monozygotic twin pairs discordant for T1D. We hypothesized that the 39 CNVs identified in the first T1D:CTRL comparison may be differentially present in this second cohort of MZ twins. For each twin, confident variant calls were catalogued in the CNP library as before. CNVs present in only 1 twin of a pair were isolated and grouped based on disease status (affected or unaffected). These variants were compared to the 39 CNVs

from the previous analysis and no overlap was found. Additionally, there were no CNVs present in more than 2 affected or unaffected twins of the pairs when this cohort was considered independently of the previous analysis.

The unaffected twin in each of these MZ twin pairs will have a greater than 75% lifetime incidence of developing islet autoantibodies and 65% of these now-unaffected twins will go on to develop T1D in their lifetime[21]. As such, CNVs may be enriched in this group as a whole that confer risk to developing islet autoimmunity or T1D. Alternatively, the CNVs depleted in the unrelated adult T1D cohort may also be depleted in the twin cohort as a whole. The 10 MZ twin pairs were compared to the CTRL cohort to determine CNVs that were enriched or depleted in the Twin group. Criteria for enrichment and depletion were identical to those outlined above. A total of 49 CNPs were enriched and 23 CNVs were depleted in the Twin cohort. Of the depleted CNVs, 22 were CNPs while 1 novel CNV was identified. All together, 72 CNVs were enriched or depleted in the Twin cohort.

The 72 CNVs identified in the Twin cohort were compared to the 39 CNVs identified in the adult T1D cohort to identify CNVs present in both cohorts. Of these, 10 CNVs were enriched in both cohorts relative to CTRL and 11 CNVs were depleted. Based upon permutation testing (with 1000 permutations), the p-value or probability of observing 10 or more overlapping CNVs in these cohorts by chance is 0.005 (= 5/1000).

The 21 CNVs were further classified to select those CNVs greater than 1,000 base pairs in length and identified by at least 3 consecutive probes on the Affymetrix array. A total of 9 CNVs (8 CNPs, 1 novel CNVR) met these criteria. Of these, 5 CNPs were enriched in the T1D and Twin cohorts and are identified by their CNP ID (Table 4-1).

These CNPs are located on 5 different chromosomes, range in size from 1,400 base pairs to 14,000 base pairs and are deletions to CNs of 0 or 1. Frequencies in the T1D and Twin cohorts range from 50% to 95% with corresponding frequencies in the CTRL cohort from 21% to 58%. CNP253, on chromosome 2p11, contains part of a non coding RNA, NCRNA00152. CNP1303 contains the gene *SNTG1*, encoding gamma syntrophin, a cytoplasmic peripheral membrane protein known to be expressed in brain. Two regions, CNP934 and CNP1162, contain at least one segment of DNA longer than 100bp with more than 70% evolutionarily conserved sequence to *Mus musculus* as determined by the ECR browser, defined as an evolutionarily conserved noncoding sequence[190]. Each of these sequences encodes at least one potential transcription factor binding site suggesting these regions may have regulatory function. The sequence encompassed by CNP1956 is not gene coding and does not contain an evolutionarily conserved noncoding sequence.

Four CNVs were depleted in the T1D and Twin cohorts relative to CTRL (Table 4-2). The 3 CNPs are gene coding regions located on 3 different chromosomes, span 3,400 base pairs to 15,900 base pairs and are also all CN deletions to 0 or 1. The frequency of the CNVs depleted in the diabetes cohorts range from total absence (0%) to 39%. The frequency of these CNPs in the CTRL cohort ranged from 42%-68%. CNP1102 contains a deletion of *TYW1*, encoding a protein involved in stabilizing ribosomal decoding processes. CNP1879 is in the region coding for the ankyrin repeat and sterile alpha motif domain gene, *ANKS1B*, and the chromosome 17 CNP2240 contains coding sequence for *TRIM16*.

Table 4-1. CNVs enriched in T1D and Twin cohorts, relative to CTRL

CNP ID[a]	Chr	Start	End	Amplification or Deletion	CTRL %	T1D %	CTRL: T1D *p*[b]	Twin %	CTRL: Twin *p*	Sequence
253	2p11	87,600,933	87,609,093	deletion	42	72	0.13	70	0.15	NCRNA00152
934	6p21	32,700,999	32,710,085	deletion	42	89	0.01	65	0.26	CNS[c]
1162	7q33	133,435,735	133,449,694	deletion	37	78	0.03	80	0.02	CNS
1303	8q11	51,194,577	51,195,974	deletion	21	61	0.03	50	0.12	SNTG1
1956	13q21	71,375,556	71,378,557	deletion	58	89	0.08	95	0.02	-

[a]CNP ID as defined in McCarroll, et al. Nature Gen 40(10):1166-74.
[b]*p*-value derived from chi-square analysis
[c]CNS= Conserved Noncoding Sequence, defined as a region >100bp with at least 70% similarity to sequence in mus musculus (as determined by ECR browser, ecrbrowser.dcode.org)

Table 4-2. CNVs depleted in T1D and Twin cohorts, relative to CTRL

CNP ID[a]	Chr	Start	End	Amplification or Deletion	CTRL %	T1D %	CTRL: T1D *p*[b]	Twin %	CTRL: Twin *p*	Sequence
1102	7q11	66,266,764	66,282,667	deletion	68	39	0.14	10	0.001	TYW1
1879	12q23	98,319,424	98,322,865	deletion	47	22	0.20	10	0.02	ANKS1B
A588[c]	15q11	18,491,920	19,803,369	both	58	33	0.24	10	0.004	BCL8, POTEB, GOLGA6L6, GOLGA8C
2240	17p12	15,483,886	15,487,515	deletion	42	22	0.35	0	0.004	TRIM16

[a]CNP ID as defined in McCarroll, et al. Nature Gen 40(10):1166-74.
[b]*p*-value derived from chi-squared analysis
[c]A588 is a novel variant identified in this study

The novel CNV, A588, depleted in both T1D and Twin cohorts is located on chromosome 15 and spans more than 1.3 million base pairs. It manifests as both an amplification and deletion and contains coding regions for genes like the golgin family members *GOLGA6L6* and *GOLGA8C*, B cell CLL/Lymphoma gene *BCL8* and an ankyrin domain family member, *POTEB*. The frequency of this variant in the CTRL group is 58%, T1D group 33% and only 10% in the Twin cohort. Interestingly, many of the variants in this region do not span the entirety of the more than 1 million base pairs (Figure 4-2), rather we see a preponderance of overlapping and non-overlapping variants clustered in this region. Because a single variant can impact regulation and expression of a gene more than 1 megabase away, these variants were grouped together as one singular CNV region (CNVR)[161].

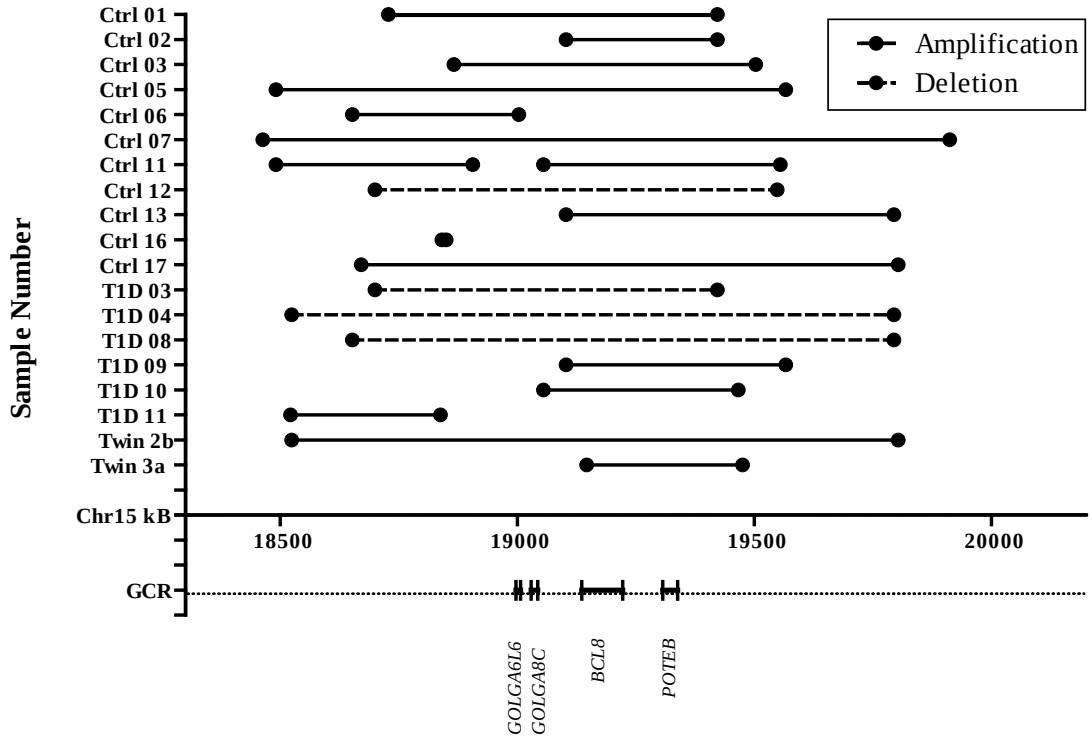

Figure 4-2. Individual breakpoints of CNVR A588.
Starting and ending points of each variant in a 1.3 mega base pair region on chromosome 15. Below the x-axis are the gene coding regions (GCR) found in this portion of the genome.

The CNVs enriched and depleted in our cohorts are potentially associated with autoimmunity so we assessed the frequency of these variants in cohorts of 21 patients with rheumatoid arthritis (RA), and 50 patients with multiple sclerosis[181] (Figure 4-3). The T1D enriched CNPs 253, 934, 1162 and 1303 also meet the criteria for enrichment in RA; additionally, CNP1162 and CNP1303 were significantly enriched in the MS cohort (Figure 4-3a). The frequency of CNP1956 did not meet the criteria for enrichment in RA or MS. For those CNVs depleted in T1D, we similarly assessed their frequency in the RA and MS cohorts (Figure 4-3b). CNPs 1879, 2240 and the novel CNVR, A588, were also

depleted in RA and MS. The depletion of CNVR A588 in RA was significant along with the depletion of CNP2240 in both RA and MS. CNP1102 was depleted in T1D and RA, but not MS (Figure 4-3b). Thus, a portion of CNVs enriched or depleted in T1D are found at similar frequencies in subjects with other autoimmune diseases.

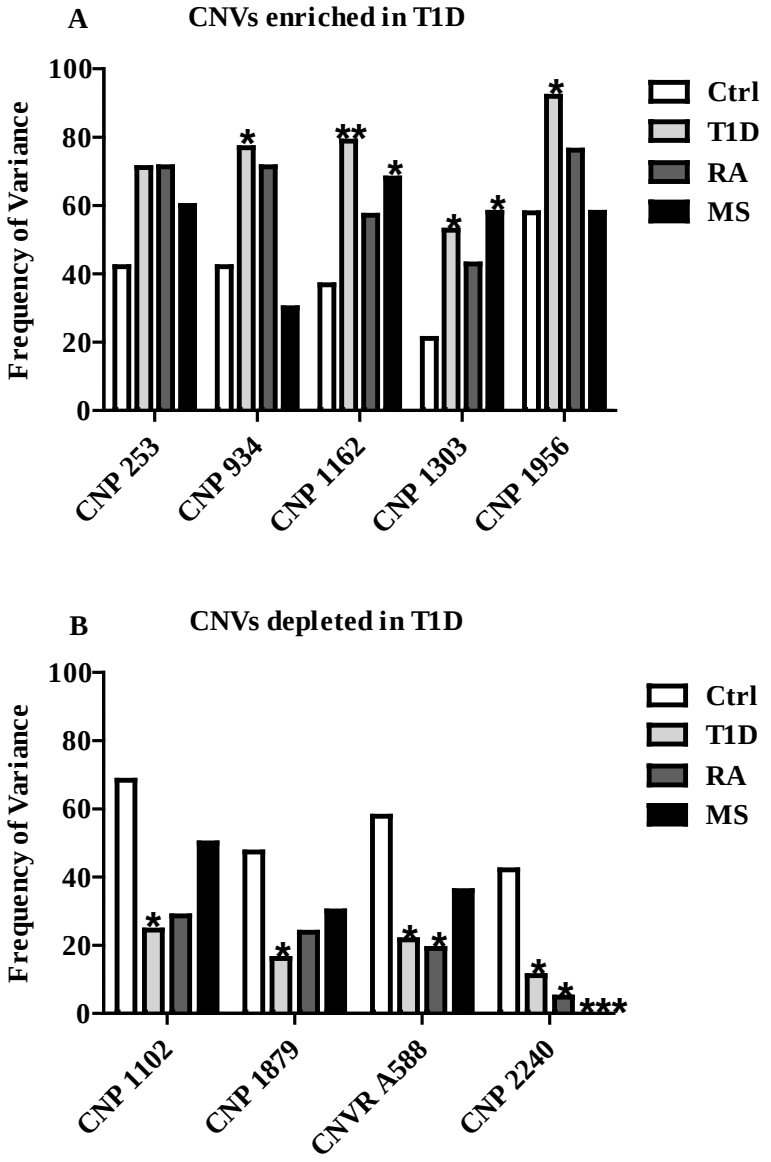

Figure 4-3. Frequencies of CNVs in other autoimmune diseases.
Panel A. Frequency of CNPs identified as enriched in the T1D cohort in the CTRL, pooled T1D, RA and MS cohorts. Panel B. Frequency of CNVs identified as depleted in the T1D cohorts, in CTRL, pooled T1D, RA and MS cohorts. CTRL n=19, T1D n=38, RA n=21, MS n=50. *p*-values determined by chi-square analysis, *= $p<0.05$, **= $p<0.005$ and ***= $p<0.0005$.

Finally, we sought to determine if similar differential frequency of variance could be seen in larger, independent cohorts of cases and controls. CN at the T1D enriched CNPs 1162, 1303 and 1956 was determined by qPCR in a group of 73 CTRL subjects and 73 subjects with T1D, independent of previous cohorts (Figure 4-4). While frequency of variance did not differ appreciably between the two groups at CNPs 1162 and 1956, the difference of variance at CNP 1303 approached significance (p=0.0655). Independent validation of 3 CNPs enriched in T1D showed one region that may continue to be of interest as a potential pathogenic variant.

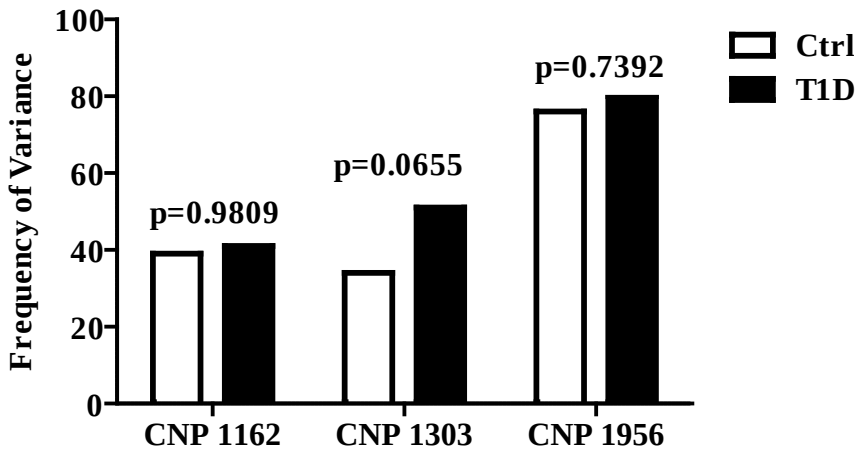

Figure 4-4. qPCR analysis of 3 T1D enriched CNPs in independent cohorts.
Frequency of variance is shown for an independent CTRL cohort (n=73) and an independent T1D cohort (n=73) at 3 CNP locations. CN was determined by ΔΔCt calculations from qPCR data. *p*-values determined by chi-square analysis.

Discussion

We identified 9 CNVs enriched or depleted in 2 independent cohorts of patients with T1D or at high risk for developing disease relative to a CTRL cohort. These CNVs represent amplifications and deletions and contain both known genes and evolutionarily conserved non-coding sequences. The regions containing these 9 CNVs were cross referenced with a list of T1D associated SNPs generated from recent reports and The National Human Genome Research Institute[26,191,192]. The only CNV region to have a corresponding SNP association is CNP934 located on chromosome 6p21, the major histocompatibility complex (MHC). The CNV region is specifically in the vicinity of HLA-DQA1; DQ alleles have long been associated with susceptibility to and protection from T1D[23]. Additionally, CNV in the HLA region has previously been reported in the literature[193]. The duplication of this finding in a small cohort of patients indicates the importance of the MHC region on chromosome 6 in the genetic susceptibility to T1D and affirms that analysis of small sample sizes can yield biologically relevant results. Additionally, because of the limited overlap of SNPs and CNVs, this study establishes the two as independent classes of genomic variants associated with T1D.

In addition to CNP934 on chromosome 6, four other CNVs were enriched in patients with T1D and unaffected twins at high risk to develop islet-specific autoimmunity and diabetes. At least 2 of the unaffected twins already test positive for islet autoantibodies. This information indicates that the 5 CNPs enriched in these groups, while they are not beacons of clinical disease, may be involved in the formation of islet autoantibodies, or autoimmunity. Enrichment or depletion of certain of these CNPs in

clinically distinct additional autoimmune diseases supports this view. CNP1303 encodes *SNTG1,* a candidate gene for scoliosis[194]. Deletions of two evolutionarily conserved non-coding sequences on chromosomes 6 and 7 could lead to dysregulation of any number of genes near the variant region on each chromosome. Underlying mechanisms by which CN deletions at these regions and others may contribute to autoimmunity remains to be determined.

Three regions enriched in T1D were further analyzed by qPCR in independent cohorts of CTRL and T1D patients, CNPs 1162, 1303 and 1956. CNPs 1162 and 1956 did not differ in their variance between the CTRL and T1D independent cohorts. CNP1303 however, is variant in a greater number of T1D subjects than CTRL and is barely shy of the criteria for enrichment with a fold change difference of 1.48. The difference in variance also approaches statistical significance. One reason the variance seen in the independent cohorts might have trended the same direction as the original data but not quite to the same degree is that our primary analysis is influenced heavily by the cohort of monozygotic twins. The affected twin of each pair was diagnosed with T1D as a child while the independent T1D cohort are adult patients with T1D who were diagnosed at varying ages. Thus, patients with an earlier onset of T1D may have a greater likelihood of possessing the CNP1303 variant.

Additionally, 4 CNVs were less likely to be variant in the T1D and Twin cohorts relative to the CTRL group. One potential consequence of these CNVs is that normal regulation and expression of these genes contributes to the T1D disease process. Alternatively, differential expression induced by variance may confer some sort of protection to the patients in the CTRL cohort.

CNVR A588 on chromosome 15 represents a unique class of variant. Both amplifications and deletions of varying lengths are observed with an increased frequency of variance in T1D. CNVs can affect expression levels of genes within 1 megabase of the variant through positional effects, by deletion or amplification of distal regulatory elements or other poorly understood mechanisms. An amplification of a promoter may cause increased expression of a certain gene. Deletion of an inhibitory element in that same region could also produce increased expression of that same gene. This T1D depleted CNVR contains as many as 26 gene coding regions and cDNA clones[195]. Preservation of this region at a CN of 2 is more common in patients with T1D for unknown reasons. Determining expression levels of the genes encoded in this region could begin to unravel the mystery of this CNVR.

In considering the overlap of enrichment and depletion of these regions in additional autoimmune cohorts of RA and MS patients we can assess contributions to autoimmunity. 2 CNPs found enriched in the diabetes cohorts were also enriched in RA and may be shared regions of susceptibility to peripheral, non-neurologic autoimmunity. An additional 2 CNPs were enriched in RA and MS and could be involved in general autoimmune processes. Similarly, the 2 CNPs and 1 novel CNVR depleted in all 3 autoimmune diseases may indicate that normal functioning in these regions is also involved in general autoimmune processes.

In conclusion, 9 CNVs were found that are either enriched or depleted in 2 independent cohorts of patients with or at high risk for developing T1D. These regions may represent genetic variants contributing to islet autoimmunity or disease onset and

could be used to assess risk of developing T1D. Knowledge of CNVs associated with T1D risk and islet autoimmunity could also improve our understanding of disease origins.

CHAPTER V

GENERAL DISCUSSION AND CONCLUSION

The framework within which we understand common complex diseases is changing. The category of risk assessment previously titled "family history" is being replaced with complex equations of SNP alleles and risk ratios. A new class of genetic variation, CNVs, is now being associated with common complex diseases and built into the aforementioned equations. Additionally, we now have the ability to assess peripheral blood gene expression of every known gene. While this technology first impacted the field of lymphocyte-mediated autoimmune diseases, in chapter II we show that peripheral blood gene expression microarray can be used to differentiate patients with diseases that are not primarily lymphocyte-mediated to detect specific types of inflammation associated with metabolic disorders.

Analysis of CNV in T1D in chapter IV yielded a number of regions of variance enriched and depleted in patients with or at high risk for developing disease. Enrichment or depletion, respectively, was also seen in the same regions in cohorts of patients with other autoimmune diseases, like RA and MS. This finding mimics those of SNP associations where 1 SNP is associated with multiple autoimmune diseases. The overlap of association of certain genetic variants, particularly SNPs, with multiple diseases, combined with the fact that seemingly none of these variants, SNPs or CNVs, are sufficient to cause disease has inspired 3 hypothesized models by which genetic variants could cause common complex disease[196].

The "common variant, multiple disease" theory of common complex disease hypothesizes that certain polymorphisms, or risk alleles, are common to a group of diseases[197]. Patients who are affected by these diseases have an overabundance of risk alleles in their respective genomes. Many of the risk alleles, however, are detectable at high frequencies in the general population. This indicates that while a given risk allele may be statistically associated with disease, each should be considered a disease "trait" rather than a causative agent. These traits, when combined with each other and potential environmental triggers, can predispose to or cause one of many related disorders.

Another model of genetic variation causing common disease is the "infinitesimal" model, originating from a study of genetic factors associated with autoimmune diseases[196]. This model states that the combination of a potentially infinite number of genomic variants, each contributing low relative risk, with environmental factors can cause a common complex disease. A study of SNP associations in T2D estimates that, assuming similar effect sizes to those SNP associations already discovered, 800 SNP variants would be needed to explain the 40% heritability of T2D[198].

The third model is called "rare alleles of major effect," or RAME[196]. In this model, variations present in extremely rare frequencies convey the genetic risk. These rare alleles might not be present at sufficient levels to meet statistical significance in a case versus control study, but may be found clustered in one individual, possibly in a homozygous fashion. The segregation of many rare alleles together, in combination with environmental risks, could cause common diseases. One example of the RAME hypothesis is the overabundance of large, rare variants in patients with autism and schizophrenia[122,184,199,200].

In the same way that variant genomic markers are thought to work in concert to comprise the heritable portions of complex diseases, we propose a model through which these markers combine with environmental exposures to impact gene expression (Figure 5-1). Altered gene expression confers a change in the cell's phenotype that ultimately results in disease.

Figure 5-1. Progression from risk to disease.
Genetic variation, currently catalogued as SNPs and CNVs, combine with environmental exposures to influence cellular gene expression. Changes in gene expression confer a change in the cell's phenotype that ultimately results in disease.

It is already possible to perform a genome-wide screen on a patient and assess for risk at many SNP alleles and regions of CNV. We hypothesize that pairing these data with continual evaluations of both environmental exposures and peripheral blood gene expression has the potential to revolutionize care and treatment of complex diseases by more accurately determining risk for disease, making it possible to arrive at an accurate diagnosis earlier in the disease process and/or make therapeutic intervention possible at the earliest stages of disease.

Thus far, risk assessment in complex diseases has primarily relied on the presence or absence of environmental risk factors (like smoking as a risk factor for chronic obstructive pulmonary disease), combined with family history and any available clinical signs (raised fasting blood glucose in T2D, for instance). Any genetic contribution to risk

is generally assessed by positive or negative family history of the disease. The nature of genetic variation is such that a majority of the risk loci are present from birth or even pre-birth. Work done to characterize potential autoimmune CNVs (Chapter IV) may add another dimension to assessment of genetic risk. Also, with peripheral blood gene expression profiling of patients with MetS, we show that a well-defined pre-disease state has a distinct gene expression profile distinguishable from CTRL subjects as well as subjects with T2D and CAD. The combination of measured genetic variants and interval screenings of environmental risk and peripheral blood gene expression profiles has the potential to produce a more accurate and fluid assessment of disease risk.

With accurate knowledge of risk, appropriate groups of patients at high risk for disease can be properly screened. Our studies of gene expression profiles associated with metabolic disorders (Chapter II) also demonstrates that portions of the MetS signature are common to both CAD and T2D while portions of these signatures are different and unique to each respective disease process. A longitudinal study of patients with MetS as they potentially progress to meet diagnostic criteria for CAD or T2D could further delineate the molecular and cellular changes that take place during the pre-clinical phase of each disease. Diagnosing a disease at the earliest point may spare patients irreversible damage to vital organs or decrease the incidence of complications most commonly found with advanced disease.

The ability to intervene and stop a disease process is only as good as its earliest detection. This point is especially relevant in T1D where it is estimated that destruction of approximately 90% of beta cells precedes the first clinical signs of hyperglycemia. Accurate assessment of disease risk prior to clinical manifestations may make it possible

to inactivate or sequester autoreactive lymphocytes before they can trigger an immune response in the pancreas, or maintain tolerance of the immune system to beta cells and prevent the autoreactive lymphocytes from becoming active.

This great promise is not without great challenge. The portrait of genetic variation is far from complete; there are serious technological and financial concerns to genotype and gene expression testing; and finally, if history is any indication, the road from a research finding to a clinical test is not one of great success.

The catalog of SNP variation has been well explored in the context of large scale studies of disease heritability[26]. GWAS of SNPs in 17,000 people, 14,000 of whom had 1 of 7 common complex diseases came to some realistic conclusions about the impact of SNPs on familiality. The authors concluded, "It is important to recognize that the association signals so far identified account for only a small proportion of overall familiality… These estimates demonstrate the limited potential of the variants thus far identified (singly or in combination) to provide clinically useful prediction of disease." And while CNV studies, ours included (Chapter IV), show that CNVs are in fact associated with a variety of distinct complex diseases, thus far there are no single CNVs with great effects on incidence and inheritance of disease, like the relationship of the amplification of the gene encoding PMP22 with Charcot-Marie Tooth disease. As GWAS of CNVs are performed and repeated, and relative risk ratios are assessed, the impact of CNVs on genetic heritability will be merged with that of SNPs into collective genomic risk alleles for any given disease.

There is an added dimension of assessing the impact of CNV both on causality and inheritance of disease and that is positional effects[201]. Knowing that a particular gene

has an increased or decreased CN does not infer a direct functional consequence. Variants may insert themselves into other regions of the genome, disrupting transcriptional regulation of genes that are neither similar in function or location to the detected CNV. Additionally, an amplified third copy of a gene may invert itself into the genome causing unknown consequences on gene regulation. Other positional effects may produce additional alterations of the transcriptional landscape that cannot be predicted simply by determining the presence or absence of a given CNV.

In addition to SNPs and CNVs, an entire class of genetic variants has yet to be fully assessed due to previous limitations in resolution of the "genome-wide" arrays. SNPs occur at the single nucleotide level and CNVs are variants larger than 1kb. Variants of intermediate size, 1 bp to 1000 bp, a class referred to as genomic "insertions and deletions", have yet to be catalogued across the genome and their possible association with complex diseases is almost completely unknown.

Technological challenges in the field of genetics range from reproducibility of array data to data interpretation and storage. With both gene expression arrays and genotyping arrays, the data are only as good as the algorithm used to interpret them. Our work in validating data derived from both gene set analysis of gene expression microarrays and Birdsuite based CNs from genotyping arrays show that data can be reproduced in a quantifiable method (Chapters II and III). For tests based on either platform to succeed in clinical disease assessment, they must be reproducible in the same sample by any person in any lab and produce an identical interpretation and clinical result.

Financially, the array alone is expensive. When one factors in the accreditation of labs to achieve technological needs, the most affordable way to put genotyping and gene expression data into practice in risk assessment and disease diagnosis is by merging multiple tests together. This is especially a concern for gene expression testing that would need to be repeated on an interval basis to serve as a marker of disease progression. If one gene expression microarray could detect transcripts differentially regulated in several common complex diseases, it would reduce the financial burden. In the realm of genetic variation, it is fathomable to place many known disease variants on one array that would be run once in a patient's life, similar to the screening for metabolic and endocrine related disorders in newborn babies. Another option is the complete sequencing of a patient's genome. While experts in the field foresee a day within years when the sequencing of a patient's genome will be "affordable," the issue of data interpretation still exists, in addition to problems with data storage as just one human genome sequence currently requires 3 gigabytes of storage, not including annotations[202].

Finally there remains the matter of taking a research finding, however sure and reproducible, and translating it to routine clinical practice. There are very few successful examples of this, with tumor typing of breast cancer arguably the most widely accepted and utilized. Assessment of risk for a complex disease using genotyping scores has been well documented in the literature but has yet to be put into practice with a schedule of screenings or preventative interventions following. While the challenges are great, gene expression and genotyping technologies have the potential to permit more accurate assessment of risk and earlier and more definitive diagnoses, which could allow earlier therapeutic interventions.

APPENDIX A

SUPPLEMENTAL DATA

Supplemental Table 2-1. Gene set analysis
Raw *p*-value for each gene set comparison between groups: RA v CTRL, MetS v CTRL, CAD v CTRL, T2D v CTRL, CAD v MetS and T2D v MetS.

Supplemental Table 2-2. Gene by gene analysis
Fold change and raw *p*-value for each gene of every gene set in each disease comparison: RA v CTRL, MetS v CTRL, CAD v CTRL, T2D v CTRL, CAD v MetS and T2D v MetS.

Supplemental Table 3-1. Array-based copy number calls
Raw CN calls for 29 samples in Partek, 29 samples in GTC and 77 samples in Birdsuite. Calls are organized by sample across the genome from chromosome 1-24.

Supplemental Table 3-2. Comparison of copy number calls
CN comparisons between Partek, GTC, Birdsuite and qPCR used to determine agreement in the tables and figures of Chapter III.

Supplemental Table 4-1. Array-based copy number calls
Raw genome CNs and confidence scores for each of the 57 patient samples in 3 cohorts (CTRL, T1D, Twin).

Supplemental Table 4-2. Birdsuite and qPCR copy number calls
Birdsuite and qPCR predicted CN calls for 185 individual comparisons.

All supplemental data can be found on the enclosed CD.

REFERENCES

[1] Marrack, P., Kappler, J. & Kotzin, B. L. Autoimmune disease: why and where it occurs. *Nature Medicine* **7**, 899-905 (2001).

[2] Ford, E. S. Prevalence of the Metabolic Syndrome Defined by the International Diabetes Federation Among Adults in the U.S. *Diabetes Care* **28**, 2745-2749 (2005).

[3] Ford, E. S., Li, C. & Sattar, N. Metabolic Syndrome and Incident Diabetes. *Diabetes care* **31**, 1898-1904 (2008).

[4] Kotzin, B. L. Systemic Lupus Erythematosus. *Cell* **85**, 303-306 (1996).

[5] Klareskog, L., Catrina, A. I. & Paget, S. Rheumatoid arthritis. *The Lancet* **373**, 659-672 (2009).

[6] Alberti, K. G. M. M., Zimmet, P. Z. & Consultation, W. H. O. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO Consultation. *Diabetic Medicine* **15**, 539-553 (1998).

[7] Bluestone, J. A., Herold, K. & Eisenbarth, G. Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature* **464**, 1293-1300 (2010).

[8] Steinman, M. D. L. Multiple Sclerosis: A Coordinated Immunological Attack against Myelin in the Central Nervous System. *Cell* **85**, 299-302 (1996).

[9] Goodnow, C. C., Sprent, J., de St Groth, B. F. & Vinuesa, C. G. Cellular and genetic mechanisms of self tolerance and autoimmunity. *Nature* **435**, 590-597 (2005).

[10] Grayson, B. in *First Aid for the Basic Sciences: General Principles* (ed TT and Krause Le, K) (McGraw-Hill, 2008).

[11] Goronzy, J. J. & Weyand, C. M. Thymic function and peripheral T-cell homeostasis in rheumatoid arthritis. *Trends in immunology* **22**, 251-255 (2001).

[12] Rocha, B., Freitas, A. & Coutinho, A. Population dynamics of T lymphocytes. Renewal rate and expansion in the peripheral lymphoid organs. *The Journal of Immunology* **131**, 2158-2164 (1983).

[13] Sprent, J., Cho, J.-H., Boyman, O. & Surh, C. D. T cell homeostasis. *Immunology & Cell Biology* **86**, 312-319 (2008).

[14] King, C., Ilic, A., Koelsch, K. & Sarvetnick, N. Homeostatic expansion of T cells during immune insufficiency generates autoimmunity. *Cell* **117**, 265-277 (2004).

[15] Kaaba, S. A. & Al-Harbi, S. A. Abnormal lymphocyte subsets in Kuwaiti patients with type-1 insulin-dependent diabetes mellitus and their first-degree relatives. *Immunology letters* **47**, 209-213 (1995).

[16] Theofilopoulos, A. N., Dummer, W. & Kono, D. H. T cell homeostasis and systemic autoimmunity. *The Journal of Clinical Investigation* **108**, 335-340 (2001).

[17] Steinman, L. Multiple sclerosis: a two-stage disease. *Nature immunology* **2**, 762-764 (2001).

[18] Silman, A. J. *et al.* Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Rheumatology* **32**, 903-907 (1993).

[19] MacGregor, A. J. *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis & Rheumatism* **43**, 30-37 (2000).

[20] Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M. & Tuomilehto, J. Genetic Liability of Type 1 Diabetes and the Onset Age Among 22,650 Young Finnish Twin Pairs: A Nationwide Follow-Up Study. *Diabetes* **52**, 1052-1055 (2003).

[21] Redondo, M. J., Jeffrey, J., Fain, P. R., Eisenbarth, G. S. & Orban, T. Concordance for Islet Autoimmunity among Monozygotic Twins. *New England Journal of Medicine* **359**, 2849-2850 (2008).

[22] Orozco, G., Rueda, B. & Martin, J. Genetic basis of rheumatoid arthritis. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* **60**, 656-662 (2006).

[23] Baisch, J. *et al.* Analysis of HLA-DQ genotypes and susceptibility in insulin-dependent diabetes mellitus. *New England Journal of Medicine* **322**, 1836-1841 (1990).

[24] Becker, K. G. *et al.* Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. *Proc Natl Acad Sci U S A* **95**, 9979-9984 (1998).

[25] Rioux, J. D. & Abbas, A. K. Paths to understanding the genetic basis of autoimmune disease. *Nature* **435**, 584-589 (2005).

[26] Wellcome Trust, C. C. C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 - 678 (2007).

[27] Ounissi-Benkalha, H. & Polychronakos, C. The molecular genetics of type 1 diabetes: new genes and emerging mechanisms. *Trends in Molecular Medicine* **14**, 268-275 (2008).

[28] Stumvoll, M., Goldstein, B. J. & van Haeften, T. W. Type 2 diabetes: principles of pathogenesis and therapy. *Lancet* **365**, 1333-1346 (2005).

[29] Kaprio, J. *et al.* Concordance for Type 1 (insulin-dependent) and Type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* **35**, 1060-1067 (1992).

[30] Ross, R. The pathogenesis of atherosclerosis: a perspective for the 1990s. *Nature* **362**, 801-809 (1993).

[31] Marenberg, M. E., Risch, N., Berkman, L. F., Floderus, B. & de Faire, U. Genetic Susceptibility to Death from Coronary Heart Disease in a Study of Twins. *NEJM* **330**, 1041-1046 (1994).

[32] Alberti, K. G. M. M., Zimmet, P. & Shaw, J. The metabolic syndrome--a new worldwide definition. *Lancet* **366**, 1059-1062 (2005).

[33] Galassi, A., Reynolds, K. & He, J. Metabolic Syndrome and Risk of Cardiovascular Disease: A Meta-Analysis. *Am J Med* **119**, 812-819 (2006).

[34] Gami, A. S. *et al.* Metabolic Syndrome and Risk of Incident Cardiovascular Events and Death: A Systematic Review and Meta-Analysis of Longitudinal Studies. *J Am Coll Cardiol* **49**, 403-414 (2007).

[35] Sattar, N. *et al.* Can metabolic syndrome usefully predict cardiovascular disease and diabetes? Outcome data from two prospective studies. *Lancet* **371**, 1927-1935 (2008).

[36] Meigs, J. B. *et al.* Body Mass Index, Metabolic Syndrome, and Risk of Type 2 Diabetes or Cardiovascular Disease. *J Clin Endocrinol Metab* **91**, 2906-2912, doi:10.1210/jc.2006-0594 (2006).

[37] Shimabukuro, M. Cardiac Adiposity and Global Cardiometabolic Risk New Concept and Clinical Implication. *Circ J* **73**, 27-34 (2009).

[38] Zimmet, P., Alberti, K. G. M. M. & Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414**, 782-787 (2001).

[39] Meigs, J. B. *et al.* Body Mass Index, Metabolic Syndrome, and Risk of Type 2 Diabetes or Cardiovascular Disease. *Journal of Clinical Endocrinology & Metabolism* **91**, 2906-2912 (2006).

[40] Eckel, R. H., Grundy, S. M. & Zimmet, P. Z. The metabolic syndrome. *Lancet* **365**, 1415-1428 (2005).

[41] Shalon, D., Smith, S. J. & Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research* **6**, 639-645 (1996).

[42] Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science (New York, N.Y* **270**, 467-470 (1995).

[43] Saeed, A. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374-378 (2003).

[44] Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25-29 (2000).

[45] Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121 (2001).

[46] Wang, L., Zhang, B., Wolfinger, R. D. & Chen, X. An Integrated Approach for the Analysis of Biological Pathways using Mixed Models. *PLoS Genetics* **4**, e1000115 (2008).

[47] Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752 (2000).

[48] Hedenfalk, I. *et al.* Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344**, 539-548 (2001).

[49] Acharya, C. R. *et al.* Gene Expression Signatures, Clinicopathological Features, and Individualized Therapy in Breast Cancer. *Journal of the American Medical Association* **299**, 1574-1587 (2008).

[50] Golub, T. R. *et al.* Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science (New York, N.Y* **286**, 531-537 (1999).

[51] Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).

[52] Bauer, J. W., Bilgic, H. & Baechler, E. C. Gene-expression profiling in rheumatic disease: tools and therapeutic potential. *Nature Reviews Rheumatology* **5**, 257-265 (2009).

[53] Olsen, N. J., Moore, J. H. & Aune, T. M. Gene expression signatures for autoimmune disease in peripheral blood mononuclear cells. *Arthritis Research & Therapy* **6**, 120-128 (2004).

[54] Maas, K. *et al.* Cutting edge: molecular portrait of human autoimmune disease. *The Journal of Immunology* **169**, 5-9 (2002).

[55] Aune, T. M., Maas, K., Moore, J. H. & Olsen, N. J. Gene expression profiles in human autoimmune disease. *Current pharmaceutical design* **9**, 1905-1917 (2003).

[56] Aune, T. M., Maas, K., Parker, J., Moore, J. H. & Olsen, N. J. Profiles of gene expression in human autoimmune disease. *Cell biochemistry and biophysics* **40**, 81-96 (2004).

[57] Maas, K., Chen, H., Shyr, Y., Olsen, N. J. & Aune, T. Shared gene expression profiles in individuals with autoimmune disease and unaffected first-degree relatives of individuals with autoimmune disease. *Human molecular genetics* **14**, 1305-1314 (2005).

[58] Liu, Z., Maas, K. & Aune, T. M. Identification of gene expression signatures in autoimmune disease without the influence of familial resemblance. *Human molecular genetics* **15**, 501-509 (2006).

[59] Maas, K., Westfall, M., Pietenpol, J., Olsen, N. J. & Aune, T. Reduced p53 in peripheral blood mononuclear cells from patients with rheumatoid arthritis is associated with loss of radiation-induced apoptosis. *Arthritis and rheumatism* **52**, 1047-1057 (2005).

[60] Kuhtreiber, W. M., Hayashi, T., Dale, E. A. & Faustman, D. L. Central role of defective apoptosis in autoimmunity. *Journal of molecular endocrinology* **31**, 373-399 (2003).

[61] Deng, X., Ljunggren-Rose, A., Maas, K. & Sriram, S. Defective ATM-p53–mediated apoptotic pathway in multiple sclerosis. *Annals of Neurology* **58**, 577-584 (2005).

[62] Ardoin, S. & Pisetsky, D. Developments in the scientific understanding of lupus. *Arthritis Research & Therapy* **10**, 218 (2008).

[63] Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc Natl Acad Sci U S A* **100**, 2610-2615 (2003).

[64] Crow, M. K. & Wohlgemuth, J. Microarray analysis of gene expression in lupus. *Arthritis Res Ther* **5**, 279-287 (2003).

[65] Yao, Y., Higgs, B., Richman, L., White, B. & Jallal, B. Use of type I interferon-inducible mRNAs as pharmacodynamic markers and potential diagnostic markers in trials with sifalimumab, an anti-IFNalpha antibody, in systemic lupus erythematosus. *Arthritis Research & Therapy* **12**, S6 (2010).

[66] Ferreira, G., Teixeira, A. & Sato, E. Atorvastatin therapy reduces interferon-regulated chemokine CXCL9 plasma levels in patients with systemic lupus erythematosus. *Lupus* **19**, 927-934 (2010).

[67] Reynier, F. *et al.* Specific gene expression signature associated with development of autoimmune type-I diabetes using whole-blood microarray analysis. *Genes and Immunity* **11**, 269-278 (2010).

[68] Klimiuk, P., Goronzy, J., Bjornsson, J., Beckenbaugh, R. & Weyand, C. Tissue cytokine patterns distinguish variants of rheumatoid synovitis. *American Journal of Pathology* **151**, 1311-1319 (1997).

[69] Aune, T. M. *et al.* Co-localization of differentially expressed genes and shared susceptibility loci in human autoimmunity. *Genet Epidemiol* **27**, 162-172 (2004).

[70] Dimas, A. S. *et al.* Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science (New York, N.Y* **325**, 1246-1250 (2009).

[71] Heap, G. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC medical genomics* **2**, 1 (2009).

[72] Lei, S.-F. *et al.* Genome-wide association study identifies two novel loci containing FLNB and SBF2 genes underlying stature variation. *Human molecular genetics* **18**, 1661-1669 (2009).

[73] Bochukova, E. G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666-670 (2010).

[74] Lettre, G. & Rioux, J. D. Autoimmune diseases: insights from genome-wide association studies. *Human molecular genetics* **17**, R116-121 (2008).

[75] International Human Genome Sequencing, C. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

[76] Cheung, V. G. & Spielman, R. S. The genetics of variation in gene expression. *Nature genetics* **32 Suppl**, 522-525 (2002).

[77] Abraham, R. *et al.* A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC medical genomics* **1**, 44 (2008).

[78] Cauchi, S. *et al.* Analysis of novel risk loci for type 2 diabetes in a general French population: the D.E.S.I.R. study. *Journal of molecular medicine (Berlin, Germany)* **86**, 341-348 (2008).

[79] Forabosco, P. *et al.* Meta-analysis of genome-wide linkage studies across autoimmune diseases. *Eur J Hum Genet* **17**, 236-243 (2008).

[80] Hokanson, J. E. *et al.* Susceptibility to type 1 diabetes is associated with ApoCIII gene haplotypes. *Diabetes* **55**, 834-838 (2006).

[81] Baranzini, S. E. The genetics of autoimmune diseases: a networked perspective. *Current Opinion in Immunology* **21**, 1-10 (2009).

[82] Barton, A. *et al.* Identification of AF4/FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes. *Hum. Mol. Genet.* **18**, 2518-2522 (2009).

[83] Coenen, M. J. H. & Gregersen, P. K. Rheumatoid arthritis: a view of the current genetic landscape. *Genes Immun* **10**, 101-111 (2008).

[84] Barker, J. M. *et al.* Prediction of Autoantibody Positivity and Progression to Type 1 Diabetes: Diabetes Autoimmunity Study in the Young (DAISY). *J Clin Endocrinol Metab* **89**, 3896-3902 (2004).

[85] Michou, L. *et al.* Linkage proof for PTPN22, a rheumatoid arthritis susceptibility gene and a human autoimmunity gene. *Proc Natl Acad Sci U S A* **104**, 1649-1654 (2007).

[86] Aarnisalo, J. *et al.* Reduced CD4+T cell activation in children with type 1 diabetes carrying the PTPN22/Lyp 620Trp variant. *Journal of Autoimmunity* **31**, 13-21 (2008).

[87] Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature genetics* **36**, 949-951 (2004).

[88] Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y* **305**, 525-528 (2004).

[89] Carvalho, C. M. B., Zhang, F. & Lupski, J. R. Genomic disorders: A window into human gene and genome evolution. *Proc Natl Acad Sci U S A* **107**, 1765-1771 (2010).

[90] Cooper, G. M., Nickerson, D. A. & Eichler, E. E. Mutational and selective effects on copy-number variants in the human genome. *Nature genetics* **39**, S22-29 (2007).

[91] Locke, D. P. *et al.* Linkage Disequilibrium and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Genome. *The American Journal of Human Genetics* **79**, 275-290 (2006).

[92] McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nature genetics* **39**, S37-42 (2007).

[93] Young, J. M. *et al.* Extensive Copy-Number Variation of the Human Olfactory Receptor Gene Family. *The American Journal of Human Genetics* **83**, 228-242 (2008).

[94] Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature reviews* **10**, 551-564 (2009).

[95] Arlt, M. F. *et al.* Replication Stress Induces Genome-wide Copy Number Changes in Human Cells that Resemble Polymorphic and Pathogenic Variants. *The American Journal of Human Genetics* **84**, 339-350 (2009).

[96] Piotrowski, A. *et al.* Somatic mosaicism for copy number variation in differentiated human tissues. *Human Mutation* **29**, 1118-1124 (2008).

[97] Feuk, L., Marshall, C. R., Wintle, R. F. & Scherer, S. W. Structural variants: changing the landscape of chromosomes and design of disease studies. *Human molecular genetics* **15**, R57-R66 (2006).

[98] Schuster-Böckler, B., Conrad, D. & Bateman, A. Dosage Sensitivity Shapes the Evolution of Copy-Number Varied Regions. *PLoS ONE* **5**, e9474 (2010).

[99] Henrichsen, C. N. *et al.* Segmental copy number variation shapes tissue transcriptomes. *Nature genetics* **41**, 424-429 (2009).

[100] Lower, K. M. *et al.* Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc Natl Acad Sci U S A* **106**, 21771-21776 (2009).

[101] Aten, E. *et al.* Methods to detect CNVs in the human genome. *Cytogenetic and Genome Research* **123**, 313-321 (2008).

[102] Carter, N. P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature genetics* **39**, S16-21 (2007).

[103] Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science (New York, N.Y* **322**, 881-888 (2008).

[104] Liang, Q., Conte, N., Skarnes, W. C. & Bradley, A. Extensive genomic copy number variation in embryonic stem cells. *Proc Natl Acad Sci U S A* **105**, 17453-17456 (2008).

[105] International, H. C. A haplotype map of the human genome. *Nature* **437**, 1299 - 1320 (2005).

[106] Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-454 (2006).

[107] Dear, P. H. Copy-number variation: the end of the human genome? *Trends in Biotechnology* **27**, 448-454 (2009).

[108] Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics* **41**, 1061-1067 (2009).

[109] Li, J. *et al.* Whole Genome Distribution and Ethnic Differentiation of Copy Number Variation in Caucasian and Asian Populations. *PLoS ONE* **4**, e7958 (2009).

[110] Lin, C.-H. *et al.* A genome-wide survey of copy number variations in Han Chinese residing in Taiwan. *Genomics* **94**, 241-246 (2009).

[111] Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics* **10**, 451-481 (2009).

[112] Lupski, J. R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66**, 219-232 (1991).

[113] Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851-855 (2006).

[114] Breunis, W. B. *et al.* Copy number variation at the FCGR locus includes FCGR3A, FCGR2C and FCGR3B but not FCGR2A and FCGR2B. *Human Mutation* **30**, E640-E650 (2009).

[115] Fanciulli, M. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics* **39**, 721-723 (2007).

[116] Mamtani, M., Anaya, J. M., He, W. & Ahuja, S. K. Association of copy number variation in the FCGR3B gene with risk of autoimmune diseases. *Genes Immun* **11**, 155-160 (2010).

[117] Nakajima, T. *et al.* Copy number variations of CCL3L1 and long-term prognosis of HIV-1 infection in asymptomatic HIV-infected Japanese with hemophilia. *Immunogenetics* **59**, 793-798 (2007).

[118] Cook, E. H., Jr. & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919-923 (2008).

[119] Glessner, J. T. *et al.* Strong synaptic transmission impact by copy number variations in schizophrenia. *Proc Natl Acad Sci U S A* **107**, 10584-10589 (2010).

[120] Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569-573 (2009).

[121] Mefford, H. C. *et al.* Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. *PLoS Genet* **6**, e1000962.

[122] Walsh, T. *et al.* Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science (New York, N.Y* **320**, 539-543 (2008).

[123] de Cid, R. *et al.* Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature genetics* **41**, 211-215 (2009).

[124] Hughes, A. E. *et al.* A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration. *Nature genetics* **38**, 1173-1177 (2006).

[125] Spencer, K. L. *et al.* Deletion of CFHR3 and CFHR1 genes in age-related macular degeneration. *Hum. Mol. Genet.* **17**, 971-977 (2008).

[126] Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).

[127] National diabetes fact sheet: general information and national estimates on diabetes in the United States, 2007. *Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention* (2008).

[128] Lloyd-Jones, D. *et al.* Heart Disease and Stroke Statistics--2009 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation* **119**, e21-181 (2009).

[129] Wellen, K. E. & Hotamisligil, G. S. Obesity-induced inflammatory changes in adipose tissue. *J Clin Invest* **112**, 1785-1788 (2003).

[130] Martinez, F. O., Gordon, S., Locati, M. & Mantovani, A. Transcriptional Profiling of the Human Monocyte-to-Macrophage Differentiation and Polarization: New Molecules and Patterns of Gene Expression. *J Immunol* **177**, 7303-7311 (2006).

[131] Tan, Q. *et al.* Differential and correlation analyses of microarray gene expression data in the CEPH Utah families. *Genomics* **92**, 94-100 (2008).

[132] Gregg, J. P. *et al.* Gene expression changes in children with autism. *Genomics* **91**, 22-29 (2008).

[133] Lockstone, H. E. *et al.* Gene expression profiling in the adult Down syndrome brain. *Genomics* **90**, 647-660 (2007).

[134] Sotiriou, C. & Pusztai, L. Gene-Expression Signatures in Breast Cancer. *New England Journal of Medicine* **360**, 790-800 (2009).

[135] Cheok, M. H. *et al.* Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet* **34**, 85-90 (2003).

[136] Holleman, A. *et al.* Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment. *New England Journal of Medicine* **351**, 533-542 (2004).

[137] Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2610-2615 (2003).

[138] Bomprezzi, R. *et al.* Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Human Molecular Genetics* **12**, 2191-2199 (2003).

[139] Arnett, F. C. *et al.* The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* **31**, 315-324 (1988).

[140] Min, J. K. & Shaw, L. J. Noninvasive Diagnostic and Prognostic Assessment of Individuals With Suspected Coronary Artery Disease: Coronary Computed Tomographic Angiography Perspective. *Circulation & Cardiovascular Imaging* **1**, 270-281 (2008).

[141] Reimers, M., Carey, V. J., Alan, K. & Brian, O. in *Methods in Enzymology* Vol. 411 119-134 (Academic Press, 2006).

[142] Wang, L., Zhang, B., Wolfinger, R. D. & Chen, X. An Integrated Approach for the Analysis of Biological Pathways using Mixed Models. *PLoS Genet* **4**, e1000115 (2008).

[143] Wang, L. *et al.* A Unified Mixed Effects Model for Gene Set Analysis of Time Course Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology* **8**, Article 47 (2009).

[144] 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 (Statistical Methodology)* **57**, 289-300 (1995).

[145] Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498-2504 (2003).

[146] Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207-210 (2002).

[147] Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367 (2009).

[148] Torkamani, A., Topol, E. J. & Schork, N. J. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* **92**, 265-272 (2008).

[149] Klareskog, L., Catrina, A. I. & Paget, S. Rheumatoid arthritis. *Lancet* **373**, 659-672 (2009).

[150] Chang, C. C. *et al.* Tolerization of dendritic cells by TS cells: the crucial role of inhibitory receptors ILT3 and ILT4. *Nat Immunol* **3**, 237-243 (2002).

[151] Shi, H. *et al.* TLR4 links innate immunity and fatty acid-induced insulin resistance. *J Clin Invest* **116**, 3015-3025 (2006).

[152] Berliner, J. A. *et al.* Atherosclerosis: Basic Mechanisms : Oxidation, Inflammation, and Genetics. *Circulation* **91**, 2488-2496 (1995).

[153] Nishimura, S. *et al.* CD8+ effector T cells contribute to macrophage recruitment and adipose tissue inflammation in obesity. *Nat Med* **15**, 914-920 (2009).

[154] Yessoufou, A., Moutairou, K. & Khan, N. A. A Model of Insulin Resistance in Mice, Born to Diabetic Pregnancy, Is Associated with Alterations of Transcription-Related Genes in Pancreas and Epididymal Adipose Tissue. *J Obes*, 654967 (2011).

[155] Hara, T., Nakayama, Y. & Gerald, L. in *Vitamins & Hormones* Vol. Volume 80 107-123 (Academic Press, 2009).

[156] Ha, E. *et al.* Interleukin 4 receptor is associated with an increase in body mass index in Koreans. *Life Sciences* **82**, 1040-1043 (2008).

[157] The International SNP Map, W. G. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933 (2001).

[158] Rudd, M. K. *et al.* Segmental duplications mediate novel, clinically relevant chromosome rearrangements. *Human molecular genetics* **18**, 2957-2962 (2009).

[159] Han, K. *et al.* L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A* **105**, 19366-19371 (2008).

[160] Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y* **315**, 848-853 (2007).

[161] Cahan, P., Li, Y., Izumi, M. & Graubert, T. A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nature genetics* **41**, 430-437 (2009).

[162] Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* **40**, 1253-1260 (2008).

[163] Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572 (2004).

[164] Fiegler, H. *et al.* Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Research* **16**, 1566-1574 (2006).

[165] Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F. & Barillot, E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413-3422 (2004).

[166] Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**, 1665-1674 (2007).

[167] Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research* **35**, 2013-2025 (2007).

[168] Dellinger, A. E. *et al.* Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Research* **38**, e105 (2010).

[169] Xu, B. *et al.* Strong association of de novo copy number mutations with sporadic schizophrenia. *Nature genetics* **40**, 880-885 (2008).

[170] Baranzini, S. E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351-1356 (2010).

[171] Bruno, D. L. *et al.* Detection of cryptic pathogenic copy number variations and constitutional loss of heterozygosity using high resolution SNP microarray analysis in 117 patients referred for cytogenetic analysis and impact on clinical practice. *Journal of Medical Genetics* **46**, 123-131 (2009).

[172] Greenway, S. C. *et al.* De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nature genetics* **41**, 931-935 (2009).

[173] Yang, T.-L. *et al.* Genome-wide Copy-Number-Variation Study Identified a Susceptibility Gene, UGT2B17, for Osteoporosis. *The American Journal of Human Genetics* **83**, 663-674 (2008).

[174] Alonso, A. *et al.* CNstream: A method for the identification and genotyping of copy number polymorphisms using Illumina microarrays. *BMC Bioinformatics* **11**, 264 (2010).

[175] Forer, L. *et al.* CONAN: copy number variation analysis software for genome-wide association studies. *BMC Bioinformatics* **11**, 318 (2010).

[176] McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics* **40**, 1166-1174 (2008).

[177] Lleo, A., Invernizzi, P., Gao, B., Podda, M. & Gershwin, M. E. Definition of human autoimmunity -- autoantibodies versus autoimmune disease. *Autoimmunity Reviews* **9**, A259-A266 (2010).

[178] Davies, J. L. *et al.* A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371**, 130-136 (1994).

[179] Tisch, R. & McDevitt, H. Insulin-Dependent Diabetes Mellitus. *Cell* **85**, 291-297 (1996).

[180] Pugliese, A. *et al.* HLA-DQB1*0602 Is Associated with Dominant Protection From Diabetes Even Among Islet Cell Antibody-Positive First-Degree Relatives of Patients with IDDM. *Diabetes* **44**, 608-613 (1995).

[181] Bruder, C. E. G. *et al.* Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. *The American Journal of Human Genetics* **82**, 763-771 (2008).

[182] Javierre, B. M. *et al.* Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Research* **20**, 170-179 (2010).

[183] Kaminsky, Z. A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nature genetics* **41**, 240-245 (2009).

[184] Grozeva, D. *et al.* Rare Copy Number Variants: A Point of Rarity in Genetic Risk for Bipolar Disorder and Schizophrenia. *Archives of General Psychiatry* **67**, 318-327 (2010).

[185] Lesch, K. P. *et al.* Genome-wide copy number variation analysis in attention-deficit/hyperactivity disorder: association with neuropeptide Y gene dosage in an extended pedigree. *Molecular Psychiatry* (2010).

[186] Willcocks, L. C. *et al.* Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. *The Journal of experimental medicine* **205**, 1573-1582 (2008).

[187] McDonald, W. I. *et al.* Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology* **50**, 121-127 (2001).

[188] Polman, C. H. *et al.* Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Annals of Neurology* **58**, 840-846 (2005).

[189] Szigeti, K., Garcia, C. A. & Lupski, J. R. Charcot-Marie-Tooth disease and related hereditary polyneuropathies: Molecular diagnostics determine aspects of medical management. *Genetics in Medicine* **8**, 86-92 (2006).

[190] Ovcharenko, I., Nobrega, M. A., Loots, G. G. & Stubbs, L. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Research* **32**, W280-286 (2004).

[191] Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics* **41**, 703-707 (2009).

[192] Hindorff LA, J. H., Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies.

[193] Wellcome Trust, C. C. C. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713-720 (2010).

[194] Bashiardes, S. *et al.* SNTG1, the gene encoding γ1-syntrophin: a candidate gene for idiopathic scoliosis. *Human Genetics* **115**, 81-89 (2004).

[195] Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996-1006 (2002).

[196] Gibson, G. Decanalization and the origin of complex disease. *Nature Reviews Genetics* **10**, 134-140 (2009).

[197] Becker, K. G. The common variants/multiple disease hypothesis of common complex genetic disorders. *Medical Hypotheses* **62**, 309-317 (2004).

[198] Pawitan, Y., Seng, K. C. & Magnusson, P. K. E. How Many Genetic Variants Remain to Be Discovered? *PLoS ONE* **4**, e7969 (2009).

[199] International Schizophrenia, C. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237-241 (2008).

[200] Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).

[201] Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).

[202] *Human Genome Project Information: Frequently Asked Questions,* <http://www.ornl.gov/sci/techresources/Human_Genome/faq/faqs1.shtml> (2010).