

**A COMPUTATIONAL ANALYSIS ON GENE FUSIONS IN HUMAN CANCERS**

By

Morgan Harrell

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

In

Biomedical Informatics

August, 2014

Nashville, Tennessee

Approved:

Zhongming Zhao, PhD

William Bush, PhD

Bing Zhang, PhD

## **ACKNOWLEDGEMENTS**

I would like to thank my advisor, Dr. Zhongming Zhao, for his guidance in this work. I would also like to thank my committee members Dr. Bing Zhang and Dr. Will Bush for their valuable feedback.

I am very grateful to all of the faculty and students in Vanderbilt's Department of Biomedical Informatics. I would especially like to thank Dr. Cynthia Gadd, the members of the Zhao lab for their time and support, and Dr. Junfeng Xia for his valuable help towards this project.

Thanks to the National Library of Medicine Training Grant 2T15LM007450-11, the Stand Up To Cancer-American Association for Cancer Research Innovative Research Grant SU2C-AACR-IRG0109, and the VICC Cancer Center Core grant P30CA68485 for supporting my education and this research.

Finally, I would like to thank my family and friends for their love and encouragement.

# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	i
TABLE OF CONTENTS .....	ii
LIST OF TABLES .....	iv
LIST OF FIGURES .....	v
CHAPTER I .....	1
Introduction.....	1
1.1. Gene fusion.....	1
1.2. Motifs and motif discovery .....	10
1.3. DNA entropy .....	14
1.4. Network analysis.....	17
CHAPTER II .....	22
Methods .....	22
2.1. Data Sources .....	22
2.2. Software and Tools .....	24
2.3. Gene fusion motif discovery .....	25
2.4. Gene fusion entropy analysis.....	27
2.5. Gene fusion network analysis .....	27

CHAPTER III.....	29
Results .....	29
3.1. Gene fusion motif discovery .....	29
3.2. Gene fusion entropy analysis.....	34
3.3. Gene fusion network analysis .....	36
CHAPTER IV.....	39
Discussion .....	39
4.1. Gene fusion motif discovery .....	39
4.2. Gene fusion entropy analysis.....	41
4.3. Gene fusion network analysis .....	42
4.4. Project Limitations .....	44
4.5. Future Work.....	45
REFERENCES .....	48
APPENDIX .....	53
Code for sliding-window entropy calculations .....	53

## LIST OF TABLES

Table 1: Summary of selected recent methods for detecting novel gene fusions .....	8
Table 2: Descriptions of the selected databases of gene fusions.....	9
Table 3: An overview of the gene fusion data deposited in COSMIC database as of 7/25/2013	22
Table 4: Details on motifs found within 600 base-pair regions of breakpoints that form gene fusions.....	30
Table 5: Details on motifs found within 300 base-pair regions of breakpoints that form gene fusions.....	32
Table 6: Over-represented transcription factor binding sites in 600 base pair regions centered at breakpoints that form gene fusions compared to control sequences .....	33
Table 7: The maximum and minimum average entropy reads for test and control sequences with p-values from t-tests.....	36
Table 8: Gene fusions with tenth percentile degrees in the true gene fusion network.....	38

## LIST OF FIGURES

Figure 1-1: An illustration of a gene fusion event. Lines represent genomic DNA and boxes represent exons. Gene A (blue) breaks (at the point designated by slanted line) and fuses with Gene B (red). Exons are numbered for clarification. Each gene has a promoter designated by the colored arrows at the beginning of the DNA. Gene B's promoter is lost and Gene A's promoter now controls transcription for the fusion.....	2
Figure 1-2: Sequence logo representing LexA binding motif. On the y-axis, "bits" is a measure of the relative base frequency via information content. ....	11
Figure 1-3: A curve showing the measure of aberrancy ( $h$ ) for element probabilities in the range 0 to 1, as calculated from equation 2.....	16
Figure 1-4: Illustrations of directed and undirected networks.....	18
Figure 1-5: Example of a random network and its degree distribution (left column) and a real (or scale-free) network and its degree distribution (right column).....	20
Figure 2-1: The distribution by chromosome position for the COSMIC reported genes participating in fusions.....	23
Figure 2-2: Percentages of COSMIC gene fusion records in given primary tissue types. ....	23
Figure 3-1: Density plot for the start positions of motifs found in the 600 base pair breakpoint sequences. The x-axis represents nucleotide position and breakpoints occur in the center at 300 base pairs. ....	31

Figure 3-2: Density plot for the start positions of motifs found in the 300 base pair breakpoint sequences. The x-axis represents nucleotide position and breakpoints occur in the center at 150 base pairs. ....33

Figure 3-3: Average entropy values across 600 base pair sequences centered at gene fusion breakpoints. Reads are in 20 base pair frames with 5 bases per shift. The red fragment of line represents reads overlapping the breakpoints. ....35

Figure 3-4: Average entropy values across 600 base pair control sequences centered at random exonic breakpoints. Reads are in 20 base pair frames with 5 bases per shift. The red fragment of line represents reads overlapping the random breakpoints.....35

Figure 3-5: Degree distribution for the 2,941 nodes in the gene fusion network (fusions and protein interactors). The distribution follows a power-law, making the network scale-free. ....36

Figure 3-6: Degree distribution for the 200 gene fusion nodes in the network. The distribution follows a power-law.....37

Figure 3-7: Box plot showing the degrees of nodes in the test gene fusion network and the control gene fusion network.....38

# CHAPTER I

## Introduction

Gene fusions are created when two or more discrete genes incorrectly join together. They are common mutations in the human genome and often found to cause cancer (Nambiar, Kari, & Raghavan, 2008). Since the advent of next generation sequencing (NGS) technology, numerous gene fusions in cancer tissues have been discovered and catalogued. We utilized the rapidly growing pool of information on gene fusions in human cancer to identify gene-fusion-formation patterns. We hypothesized that understanding the mechanisms and risk for gene-fusion formation could facilitate novel gene fusion detection in human cancers and lead to better targeted treatments. This thesis project has three related computational analyses: 1) we used a motif discovery tool to examine common sequence patterns at and around breakpoints that form fusions, 2) we calculated entropy in a sliding-window manner to determine structural characteristics at and around breakpoints that form fusions, and 3) we executed a gene-fusion network analysis to visualize and compare cancer-associated gene fusion metrics versus controls.

### 1.1. Gene fusion

A gene fusion is the result of two or more discrete genes incorrectly joining. These mutations may result in fused mRNA transcripts. In most two-gene gene fusion events, the promoter from the gene in the 5' position controls transcription. Gene fusions result from



translocations, inversions, insertions or deletions and are common mutations. Nambiar et al. (Nambiar et al., 2008) reports that approximately 20% of cancers are driven by gene fusions caused by chromosomal locations. Figure 1-1 illustrates a gene fusion event. Because of the estimated large percentage of cancers caused by gene fusions, the research community has great interest in gene fusions that facilitate tumorigenesis. Identifying and determining mechanisms by which gene fusions act lead to projections toward drug targets for cancer treatment.

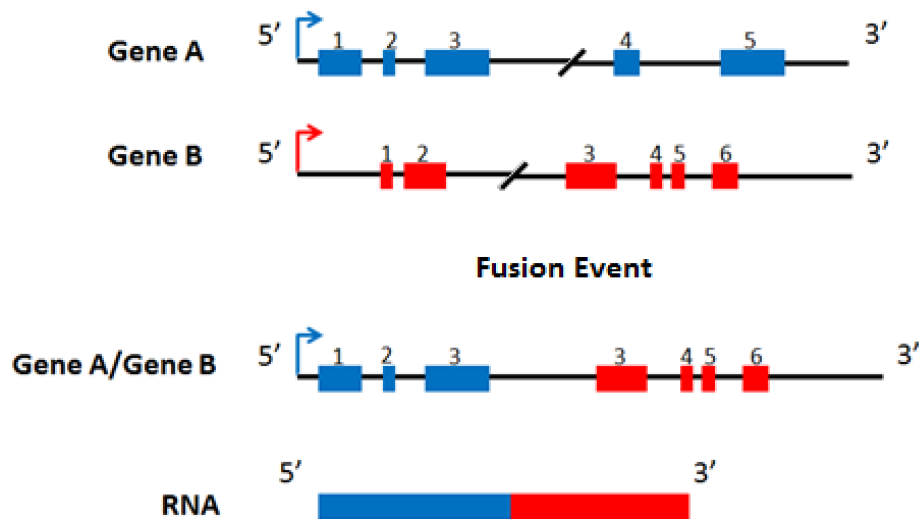


Figure 1-1: An illustration of a gene fusion event. Lines represent genomic DNA and boxes represent exons. Gene A (blue) breaks (at the point designated by slanted line) and fuses with Gene B (red). Exons are numbered for clarification. Each gene has a promoter designated by the colored arrows at the beginning of the DNA. Gene B's promoter is lost and Gene A's promoter now controls transcription for the fusion.

One of the most successful gene fusion discoveries involves BCR-ABL in chronic myelogenous leukemia (CML). As early as 1960, Nowell and Hungerford examined leukemia cells from chronic phase CML patients and found a common abnormality - in each of their seven

samples, chromosomes 9 and 22 contained reciprocal translocations (P. C. Nowell, 1960). Their finding ignited additional studies that began to answer why this abnormality, which they called the Philadelphia Chromosome, is associated with CML. Two genes are disrupted in a Philadelphia chromosome. Abelson murine leukemia viral oncogene (*ABL*), translocated from chromosome 9, is a proto-oncogene that encodes a tyrosine kinase and is active during cell differentiation, cell division, cell adhesion, and stress response. Breakpoint cluster region gene (*BCR*), from chromosome 22, encodes a protein with serine/threonine kinase activity but is otherwise uncharacterized at this time. Not only does a BCR-ABL gene fusion occur on the Philadelphia chromosome, but a novel mRNA transcript is produced that contains components from both genes (Ben-Neriah, Daley, Mes-Masson, Witte, & Baltimore, 1986).

Once the structural mutation on the Philadelphia chromosome was characterized, researchers began answering questions about the relationship between the gene fusion and CML. The fusion protein BCR-ABL behaves as an abnormal tyrosine kinase that over-phosphorylates substrates and facilitates tumorigenesis (reviewed in (Wong & Witte, 2004)). ABL tyrosine kinase activation is in part regulated by a SH3 domain which is lost during the gene fusion event, thus turning ABL into an oncogene (Barilá & Superti-Furga, 1998). The role BCR plays in the gene fusion remains unclear. There is evidence that ABL oligomerization is sufficient for tumorigenesis but in fusions with different BCR breakpoints or with other gene partners, ABL tyrosine kinase activity levels change (reviewed in (Wong & Witte, 2004)).

Research identifying the BCR/ABL gene fusion and understanding how it leads to tumorigenesis culminated into developing imatinib, a drug with success treating patients with

CML. Imatinib is a tyrosine-kinase inhibitor that prevents fusion protein BCR-ABL activity, thus preventing cancer cell growth (reviewed in (Wong & Witte, 2004)). Since BCR-ABL is predominantly expressed cancer cells, imatinib is classified as a targeted therapy. A randomized study on 1106 CML patients ended with 96.7% of imatinib-treated patients, compared to 91.5% of alternatively-treated patients, classified as “free from CML progression” (O’Brien et al., 2003). The drug passed clinical trials and is on the market as first-line treatment since 2001 (<http://www.accessdata.fda.gov>).

BCR-ABL is only one of the tyrosine kinase gene fusions involved with malignancies. Identified from epithelial tumors, there are currently ten tyrosine kinase genes that form fusions with at least one partner gene (reviewed in (Shaw, Hsu, Awad, & Engelman, 2013)). The mechanisms by which tyrosine kinase fusions lead to tumorigenesis are similar across different types of cancer – they involve deregulation of tyrosine kinase activity. For example, receptor tyrosine kinase and proto-oncogene RET (abbreviation of “rearranged during transfection”) forms fusions with several partners in papillary thyroid cancer. These gene fusions contain RET’s intact tyrosine kinase domain fused to the active promoter of the partner gene, which ultimately leads to incessant MAPK (mitogen-activated protein kinase) signaling and tumorigenesis in thyroid cells (reviewed in (Nikiforov & Nikiforova, 2011)).

Aberrant tyrosine kinase activity is not the only mechanism by which gene fusions facilitate tumorigenesis. A class of gene fusions common in prostate cancer consist of an androgen-controlled genomic regulatory element fused to an oncogenic transcription factor in the ETS (E26 transformation specific, where E26 is leukemia virus) family (Kumar-Sinha,

Tomlins, & Chinnaiyan, 2008). The five prime partners in these gene fusions (for example, TMPRSS2) are prostate-specific and have ubiquitous expression. Fusing their promoters onto an ETS family member leads to overexpression of oncogenic transcription factors, which in turn leads to cellular pathway alterations and loss of tumor suppressor activity (reviewed in (Kumar-Sinha et al., 2008)).

There are two main mechanisms by which gene fusions can facilitate cancer: a fusion gene can yield a chimeric protein with aberrant activity (e.g. BCR-ABL), or a promoter or enhancer can fuse to a proto-oncogene and lead to overexpression of an oncogenic protein (e.g. TMPRSS2 fusions) (Aman, 1999). As stated previously, not all gene fusions are pathogenic. Cancer is a genetic disease in which cells accrue many mutations – some of which actively participate in disease progression (called drivers), and some of which result from disease progression (called passengers). Identifying gene fusions and differentiating drivers from passengers are both current challenges in gene fusion research.

NGS is a sequencing strategy that yields high-throughput and low-cost sequence data. NGS augments the amount of available whole genome, exome, and transcriptome sequence data, including data from cancer tissues with gene fusions. The number of reported gene fusions from tumor tissue has grown exponentially during the last six years. Mitelman et al. (Mitelman, Johansson, & Mertens, 2004) estimated that the number of fusion genes in a cell is a linear function of the number of chromosomal aberrations and Futreal et al. (Futreal et al., 2004) reported that the most common mutations in cancer are translocations that create chimeric genes. Nevertheless, generating raw sequence data does not guarantee gene fusion

identification. Detecting novel gene fusions in cancer enhances the current catalog, and detecting targeted gene fusions facilitates classifying cancer types in patients. There are targeted and unguided methods for detecting gene fusions in cancer tissue.

Targeted gene fusion identification is a necessary tool for translating gene fusion research to clinical treatment. The type of cancer dictates cancer treatment - it is advantageous to prescribe a tyrosine kinase inhibitor to a patient with abnormal-tyrosine-kinase-driven cancer, but that treatment will be ineffective if the cancer stems from a different cause. To help determine treatment, detecting the presence of a gene fusion driver from tumor tissue is necessary. Exploring regions surrounding breakpoints highlights targets for directed gene fusion identification. For example, specific sequence patterns may indicate gene fusion presence. This is the drive for motif discovery on sets of gene fusions (further discussed in section 1.2). Chmielecki et al. (Chmielecki et al., 2010) discovered that GXGXXG motifs exist near tyrosine kinase fusions, and they performed targeted next generation sequencing around those motifs to identify fusions. Targeted sequencing could identify suspected gene fusions from patient samples and obviate extraneous sequencing. Another example where targeted gene fusion detection assists in patient diagnostics involves ETS family gene fusions. ETS family gene fusions act as biomarkers for prostate cancer and exist in up to 70% of the cases (Thieme & Groth, 2013). The company KREATECH (<http://www.kreatech.com>) produces kits to detect ETS family gene fusions with the aim of diagnosing prostate cancer.

Novel gene fusion identification is a tough problem to tackle and several research groups have published methods to detect fusions from RNA-Seq and DNA-Seq data. Solutions

involve constructing an algorithm that can align sequences and highlight anomalies suggesting gene fusion events, such as exons from discrete genes present on the same sequence read. Trans-splicing, intergenic splicing, and a plethora of genomic mutations in tumor cells augments type I and type II errors in fusion-event screening against background sequences. To date, there is no robustly tested and error-free algorithm that identifies novel gene fusions from sequencing data. Table 1 lists five recent algorithms for gene fusion detection, and includes a brief summary of their results. The algorithms have different input sequence requirements and limitations, and have each identified at least one novel gene fusion when evaluated.

The number of reported gene fusions increases, and it befits researchers to have access to descriptive information on them. Public databases such as Mitelman and COSMIC hold records on gene fusions for research aid. Information on the gene fusions includes, at minimum, the name and position (5'/3') of genes participating in the fusion. Additional information can consist of sequence information, the type of cancer in which the fusion was found, the reference paper, etc. Table 2 lists four of the current databases that curate information on gene fusions.

**Table 1: Summary of selected recent methods for detecting novel gene fusions**

<b>Tool</b>	<b>Input</b>	<b>Technique/advantages</b>	<b>Limitations</b>	<b>Reference</b>
<b>deFuse</b>	RNA-Seq data	Considers all alignments rather than best-fit alignments. Considers all locations for fusion boundaries rather than focus at ends of known exons.	Requires at least five discordant read pairs to detect a gene fusion, which leads to missed fusions with low-expression.	McPherson et al., 2011
<b>Genome Fusion Detection</b>	SNP-array data	Detects genes at the transition regions where copy number variations occur. Can detect non-functional, silenced and novel fusions	Detects fusion genes from only unbalanced mutation events.	Thieme & Groth, 2013
<b>FusionMap</b>	RNA-seq data	Aligns fusion reads to the genome without prior knowledge of probable fusion regions. Can detect fusions from single or paired-end reads.	Relies on long read lengths and requires longer computational time compared to other methods.	Ge et al., 2011
<b>FusionQ</b>	Paired-end RNA-seq data	In addition to detecting gene fusions, this tool constructs the chimerical transcript structures and estimates their quantity.	Reports less gene fusions compared to other methods due to increased filter constrictions.	Liu, Ma, Chang, & Zhou, 2013
<b>FusionSeq</b>	Paired-end RNA-seq data	Ranks gene fusion candidates by several statistics and identifies sequence at exact breakpoints.	Fusion detection depends on a gene annotation set for information on genes and their isoforms, so candidate fusion identification is limited to the set.	Sboner et al., 2010

**Table 2: Descriptions of the selected databases of gene fusions**

<b>Database</b>	<b>Content</b>	<b>Number of gene fusion records</b>	<b>Website</b>	<b>Reference</b>
<b>ChimerDB</b>	Fusion transcripts collected from public resources	11,747	<a href="http://biome.ewha.ac.kr:8080/FusionGene/">http://biome.ewha.ac.kr:8080/FusionGene/</a>	Kim et al., 2010
<b>COSMIC</b>	Gene fusions, genomic rearrangements, and copy number variations	9,054	<a href="http://cancer.sanger.ac.uk/">cancer.sanger.ac.uk/</a>	Forbes et al., 2010
<b>Mitelman</b>	Chromosome aberrations and gene fusions	2,038	<a href="http://cgap.nci.nih.gov/Chromosomes/Mitelman">http://cgap.nci.nih.gov/Chromosomes/Mitelman</a>	Mitelman F, Johansson B, Mertens F, 2014
<b>TICdb</b>	Translocation breakpoints in cancer	1,374	<a href="http://www.unav.es/genetica/TICdb/">http://www.unav.es/genetica/TICdb/</a>	Novo, de Mendíbil, & Vizmanos, 2007

Gene fusions are classified as drivers or passengers – drivers play an active role in facilitating tumorigenesis while passengers do not. The pool of known gene fusions grows, and there is strong demand to identify the gene fusions drivers for targeted studies. Research groups put forth several approaches in attempt to accomplish this. Wang et al. (Wang et al., 2009) published a method to predict a gene fusion’s driver likelihood with a ConSig score. The ConSig score, derived from gene ontologies, assumes the driver probability for a gene is correlated with the gene’s association to oncogenic-related gene ontologies. The authors report with low power that cancer-related fusion genes share common gene ontologies. Shugay et al (Shugay, Ortiz de Mendíbil, Vizmanos, & Novo, 2013) developed *Oncogene*, an algorithm to predict oncogenic potential of gene fusions based on known features in gene fusion drivers. They focused on sequences including fusion protein domains and protein interaction interfaces



to rank driver potential. They had positive results with validation tests, but they depend on the current set of validated gene fusion drivers. Despite the algorithm development toward classifying gene fusion drivers, we lack a robustly tested method with positive results.

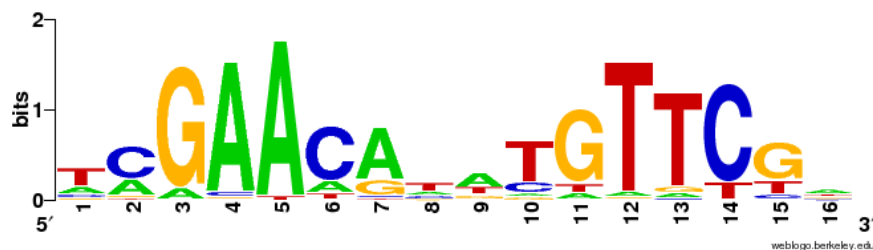
Gene fusions drive an estimated 20% of human cancer (Nambiar et al., 2008) and are of great interest in cancer research and treatment. There is clinical success from translating BCR-ABL research into a usable drug and in diagnosing prostate cancer via targeted gene fusion detection. There is room for improvement in methods for detecting gene fusions and differentiating drivers from passengers. Additionally, there are knowledge gaps in what the scientific community understands about gene fusions. It is mostly unclear how gene fusions form. Radiation plays a role in gene fusion formation in papillary thyroid carcinomas (Mizuno, Kyoizumi, Suzuki, Iwamoto, & Seyama, 1997), and a handful of sequences common at breakpoints in leukemia and lymphoma are reported (reviewed in (Aman, 1999)). A comprehensive study on where gene fusions form is lacking. Gene fusion research is in a positive direction and, with enhanced sequencing technologies, continues to move forward.

## **1.2. Motifs and motif discovery**

Motifs are short conserved sequences associated with biological functions. Among other roles, motifs denote binding sites or splice junctions in DNA, and active sites or domains in proteins. Type II restriction enzymes, which cut DNA in bacteria, must bind to specific motifs. *EcoRI* binds to 6-mer 5'-GAATTC-3' and incorrect binding leads to detrimental DNA shearing (Pingoud & Jeltsch, 2001). Motifs become more complicated with degenerative properties. *Escherichia coli* require two conserved regions for transcription initiation: a "TATA box" 10 base

pairs (bps) upstream and a TTGACA motif 35 bps downstream from the start site. Actual transcription initiation sites match 54-82% (7-9 out of 12 base pairs) of these motifs (reviewed in (D’haeseleer, 2006)). Despite exchanged base pairs, transcription executes proving that motifs do not always need to be exact sequences to be functional.

The scientific community uses sequence logos to represent motifs with degenerative properties. A sequence logo is a position-dependent symbol-probability matrix – it depicts the probability of possible symbols at each position in a graph. Figure 1-2 shows a sequence logo for repressor enzyme LexA’s binding motif. The y-axis on the sequence logo represents “bits,” which are binary digits with the information required to distinguish between two possibilities. Two bits are required for the four nucleotides A, T, C and G (00, 01, 10, and 11). Bits measure the information content at a given position (Section 1.3 explains information content in detail). A site with no sequence conservation will have 0 bits of information content, or an equal chance of having A, T, C, or G. A completely conserved site will have 2 bits of information content. The total information content at a position follows the equation  $2 - \text{the uncertainty of that position}$ , and the height of a symbol at that position is the information content multiplied by the nucleotide’s relative frequency.



**Figure 1-2: Sequence logo representing LexA binding motif. On the y-axis, “bits” is a measure of the relative base frequency via information content.**

When an n-mer repeats itself across polymer sequences, there is evidence that it does not occur by chance and may be a motif. This is the basis for motif discovery. Identifying a known motif in a sequence gives good indication on what biochemical process occurs at that region. Identifying common motifs among sequences evinces that there are mutual biological mechanisms acting at those regions. Motif discovery among sequences is a computational process and there are three different categories of motif discovery algorithms: enumeration, probabilistic optimization, and deterministic optimization (D'haeseleer, 2006).

Enumerative motif discovery algorithms use exhaustive search on all possible motifs in a search space (Sinha & Tompa, 2002). An enumerative algorithm keeps a running total of n-mer occurrence in the target sequences, then reports those with the highest totals. Since enumerative motif search is so thorough, these algorithms are not limited to a local optimum. Nevertheless, restricting searches to exact nucleotide sequences may lead to overlooked actual binding sites, which are often degenerate. These algorithms may miss flexible motifs.

Probabilistic optimization (or stochastic) algorithms depend on Gibbs sampling for motif discovery. Gibbs sampling is sampling by one variable at a time conditioned on all other variables. Probabilistic optimization algorithms work by sampling for motif start points, building a position specific scoring matrix for all points excluding one, then setting the excluded point to the position that best matches the matrix model (Thijs et al., 2002). The probability that a symbol exists at a given position serves as a measure for sequence scores. Scores converge until alignment no longer changes, resulting in a projected motif.

Deterministic optimization algorithms depend on expectation maximization for motif discovery. These algorithms iterate through two steps: the expectation step where current parameters project an “expected” structure, and the maximization step where the expected structure feeds back to re-estimate parameters (Dempster, 1977). Over iterations, the initial structure converges to one with maximum log likelihood in the search space, and results in a final motif. Expectation maximization is a local optimization – the final motif is sensitive to the initial expected structure. Therefore, it does not guarantee the model with global maximum log likelihood for the search space (Blekas, Fotiadis, & Likas, 2003). For optimization, one can repeat these algorithms under different initial parameters to better chances of finding motifs.

Sequence data quality limits performance for each of these motif discovery methods. For best results, one should restrict data to high-quality sequences. It is important to consider nucleotide background frequencies in the motif search space. Otherwise, most motif-discovery tools will assume equal probabilities for all nucleotides, skewing the motif’s statistical significance measurements.

If motifs existed near gene fusion regions, they would help explain where and why gene fusions happen, and improve the detection of gene fusions from complex data generated by NGS. Gene fusion-associated motifs may act as markers for fragile DNA regions or suggest implication of a binding factor on gene fusion formation. Myers et al. (Myers, Freeman, Auton, Donnelly, & McVean, 2008) identified two sequence motifs associated with recombination hot spots and genome instability: a 7-mer CCTCCCT and 13-mer CCNCCNTNCCNC (“N” denotes any base type). If these motifs are enriched near gene fusion breakpoints in cancers, we have

evidence to predict cancer-associated gene fusion formation results from aberrant crossover events. Chmielecki et al. (Chmielecki et al., 2010) found that tyrosine-kinase gene fusion breakpoints occur upstream from a GXGXXG amino acid motif (“G” signifies glycine and “X” signifies any amino acid). Finding a motif when limited to tyrosine-kinase gene fusions supports that gene-fusion motifs are probable and performing motif discovery across many cancer-associated gene fusions may yield significant information.

After motif discovery and function identification, motifs are curated in public databases like TRANSFAC (Matys et al., 2006) and JASPAR (Bryne et al., 2008). TRANSFAC (TRANSCRIPTION FACTOR database) contains information on eukaryotic transcription factors with their respective genomic binding sites and DNA binding profiles. The database offers tools to facilitate motif based research. *Match* is a weight matrix-based program that predicts transcription factor binding sites from their database in DNA sequences. *F-Match* builds off *Match* and identifies statistically over-represented transcription factor binding sites in a set of sequences compared against a control sequence set. If a known binding factor is implicated in gene-fusion formation, it will have enriched binding sites near breakpoints that form fusions. *F-Match* is an appropriate tool to identify those binding sites and evince differences compared to control sequences.

### **1.3. DNA entropy**

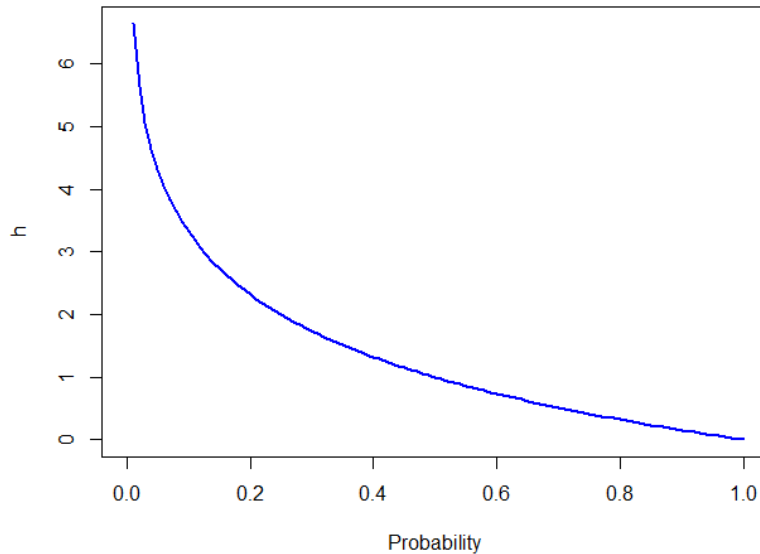
Entropy measures the number of ways a thermodynamic system can be arranged and is commonly thought of as a gauge for disorder and uncertainty. Entropy is inversely correlated with information content, making information content a measure of certainty. The base of these measurements requires determining the number of binary decisions required to ascertain

information. The formula for determining the number of binary decisions ( $n_q$ ) for a set of  $N$  options is:

$$n_q = \log_2 N \quad (1)$$

As a simple example, imagine the set for four nucleotides {A, C, T, G} where one nucleotide is selected and a person must determine which nucleotide it is using binary questions. The person may probe if the nucleotide is in the set {A, C}, then further reduce the set into single nucleotides.  $\log_2 4 = 2$  questions are needed to determine the nucleotide. This equation assumes the probability for all elements of the set are equal. When elements have disparate probabilities, we can gauge aberrancy for an element's occurrence with the expanded equation below, where  $h_i$  signifies aberrancy from seeing that element, and  $p_i$  signifies the element's probability for occurrence (Schneider, Stormo, Gold, & Ehrenfeucht, 1986). Figure 1-3 shows the aberrancy curve  $h$  for element probabilities in the range 0 to 1.

$$h_i = -\log_2 p_i \quad (2)$$



**Figure 1-3: A curve showing the measure of aberrancy ( $h$ ) for element probabilities in the range 0 to 1, as calculated from equation 2.**

Claude Shannon coined the “uncertainty measure,” or Shannon Entropy, which is the average aberrancies weighted by their occurrence probability. Shannon entropy is calculated by the equation below (C. E. Shannon, 2001).

$$H = \sum_i p_i h_i = -\sum_i p_i \log_2 p_i \quad (3)$$

Schneider defines information content, based on Shannon’s uncertainty, as the difference in global uncertainty and uncertainty at a given position ( $j$ ), calculated by the equation below (Schneider et al., 1986).

$$IC_j = H_g - H_j \quad (4)$$

Entropy estimates on DNA provide information about genomic complexity and arrangement. Fragments of DNA under selective pressure have lower entropy than those that

are not – exons have lower entropy than introns and older fragments of DNA have lower entropy than younger (Koslicki, 2011). Farach et al. used entropy measurements as a way to detect DNA splice junctions (Farach et al., 1995). A second way to think of Shannon entropy is a measure for how efficiently data could be compressed without loss. Higher the entropy means less redundancy and lesser compression. Gatlin proposes that high redundancy DNA molecules with optimum composition may have lower probability of error, therefore it is appropriate that exons and conserved regions have lower entropy (Gatlin, 1968).

A sliding window entropy analysis reveals changes in complexity throughout a DNA sequence. The sliding window works by setting a frame length and start position on a DNA sequence. The region within the frame yields a Shannon entropy measure, which is recorded, then the frame shifts a determined number of base pairs downstream and another read is taken. Once graphed, the entropy measures show changes, oscillations, and conserved regions that describe the sequence. Performed at gene fusion breakpoints, we can gauge patterns in entropy that may help explain where gene fusions happen and why. We hypothesize an entropy increase at breakpoint regions that signifies higher probability for error.

#### **1.4. Network analysis**

A network is a graphical representation of data that consists of nodes, which represent entities, and edges, which represent relationships. A network can be directed or undirected. Directed networks have edges that are one way denoting  $X$  acts on  $Y$  but  $Y$  does not act on  $X$ . An example of a directed network is a food chain – a fox eats the rabbit and the rabbit does not eat the fox. Undirected networks denote mutual relationships. An example of an undirected



network is a social network – Anne is friends with Beth therefore Beth is friends with Anne. To differentiate between the two graphically, directed networks have arrows for edges while undirected networks have lines Figure 1-4.

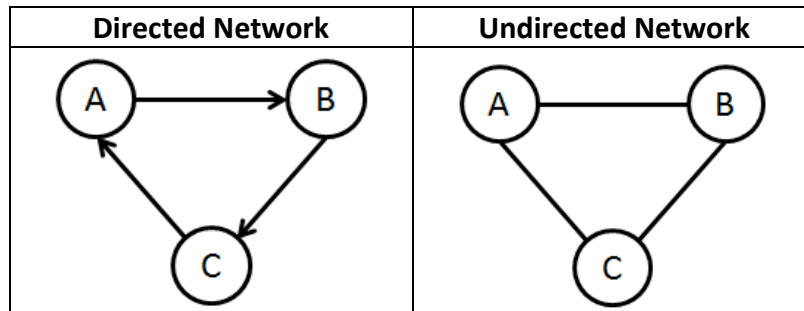


Figure 1-4: Illustrations of directed and undirected networks.

A mathematical descriptor of nodes is centrality, a measure of weight or importance in a network. There are three categories of centrality: degree centrality, closeness centrality, and betweenness centrality.

Degree centrality is based on the node’s degree, which is the number of connections, or edges, it has to other nodes. In directed networks, degree is categorized by in-degree and out-degree (the number of edges directed away from the node and the number of edges directed toward it. A node’s degree centrality can be interpreted as the probability of intercepting that node when moving through the network.

Closeness centrality is a measure for how easily one can move to the rest of the network from the given node. The lowest number of edges that need to be traversed to reach one node from another defines their distance. The “farness” of a node is the sum of its distances to all of the other nodes, and its “closeness” is the inverse of the “farness” (Sabidussi, 1966).

Betweenness centrality is the total number of times a node acts as a bridge along the shortest path between two other nodes. It is interpreted as a measure for control of movement through the network (Freeman, 1977). Nodes with high betweenness have regulatory power over information flow through the network.

Networks are classified by their degree distribution (Figure 1-5). Random networks manifest a normal degree distribution – the degrees of all nodes are distributed around and average. Scale-free networks manifest a power-law distribution in which most nodes have low degrees and a few nodes have high degrees. When observing any sub-region of the network, the degree distribution will always follow the power-law, hence the name scale-free. Biological networks, such as those that depict gene regulation and signaling are scale-free (Albert, 2005). Höglund et al. built the first gene fusion network where nodes represent participating genes and edges represent a gene fusion event. They reported that the network's organization demonstrates a scale-free network topology with the power law degree distribution found in naturally occurring networks (Höglund, Frigyesi, & Mitelman, 2006).

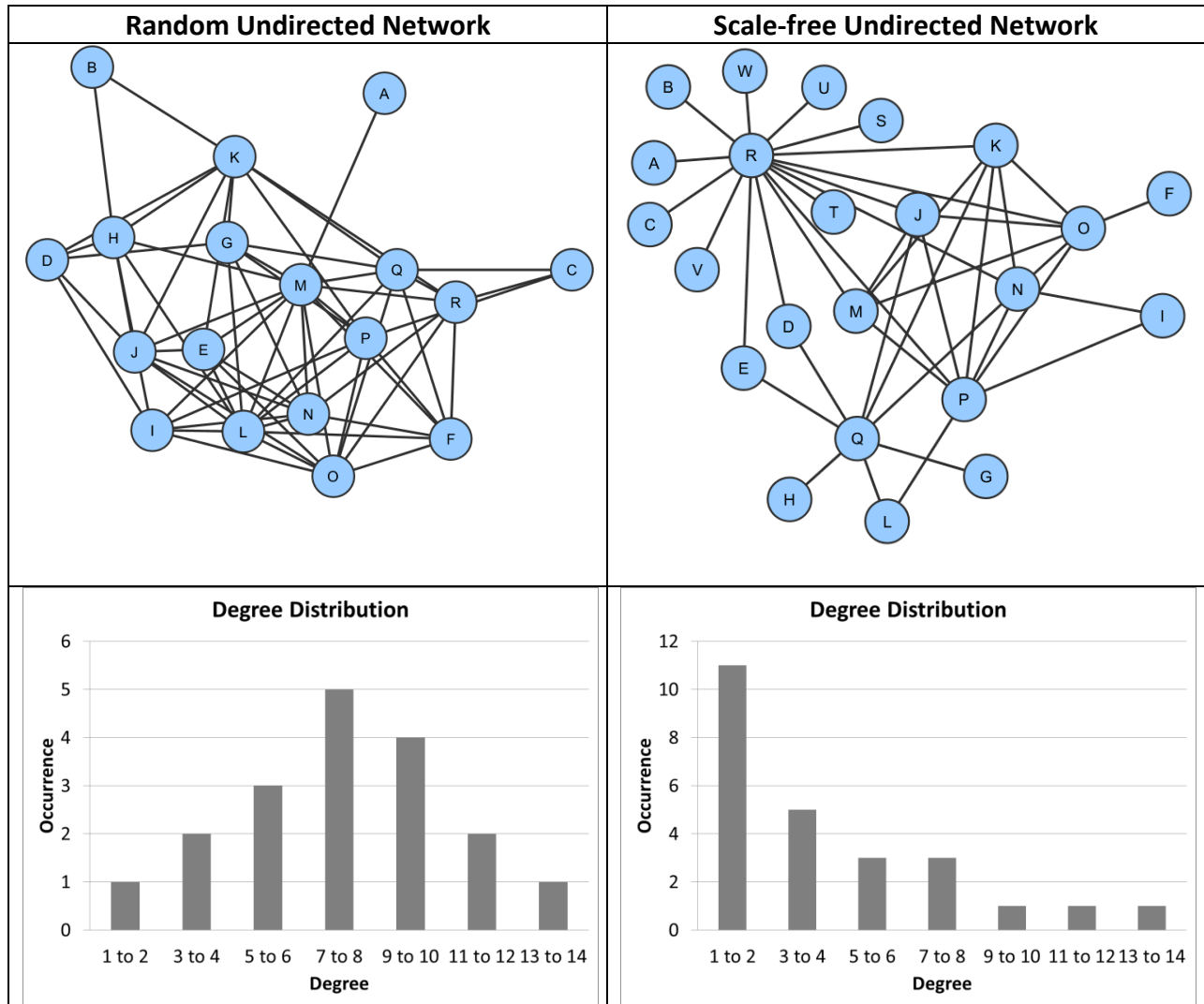


Figure 1-5: Example of a random network and its degree distribution (left column) and a real (or scale-free) network and its degree distribution (right column).

Protein-protein interaction (PPI) networks reflect functional interactions such as cellular pathways – nodes represent proteins and edges denote interactions with other proteins. A PPI network depicts which proteins interact, which protein groups are modular, and where feedback loops occur. Wu et al. (Wu, Kannan, Lin, Yen, & Milosavljevic, 2013) mapped gene fusions to a PPI network for the purpose of highlighting gene fusion drivers. They hypothesized that a gene fusion is more likely to be a driver if one or more participating genes act as a hub in

a PPI network. The more connected a gene is, the more deregulation it would cause if disrupted. Their study used data on known cancer fusions from the Cancer Gene Census database (Futreal et al., 2004) and their method correctly predicted most of the 38 fusions with oncogenic importance in their test set (with 19% false positive rate).

We chose to explore the role gene fusions have in cancer and their impact on cellular pathways with network analysis. We tested the hypothesis that gene fusions in cancer are more disruptive compared to passive mutations by comparing network metrics across true gene fusions and randomly generated gene fusions. We used two-gene gene fusion data from COSMIC and randomly generated gene fusion sets. After mapping fusion genes in the PPI network, we compared the control and real-data gene fusion's degree measurements.

## CHAPTER II

### Methods

#### 2.1. Data Sources

All gene fusion records were collected from the Catalog of Somatic Mutations in Cancer (COSMIC) (<http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/download>) (Forbes et al., 2010). COSMIC is a public database for information on gene fusions, genomic rearrangements, and copy number variations in human cancers. The database curates mutations reported in scientific literature, and mutations reported from the Cancer Genome Project (CGP) at the Sanger Institute UK. Table 3 shows general statistics on COSMIC's gene fusion dataset as of the date 7/25/2013. Figure 0-1 shows the chromosomal distribution of genes participating in fusions for our dataset, and Figure 0-2 shows the distribution by the primary tissue type that the gene fusions were reported from.

Table 3: An overview of the gene fusion data deposited in COSMIC database as of 7/25/2013

<b>Number of reported gene fusions</b>	9054
<b>Number of gene fusions with sequence data</b>	662
<b>Number of participating genes</b>	222
<b>Number of unique gene pairs</b>	200
<b>Number of unique gene fusion sequences</b>	462

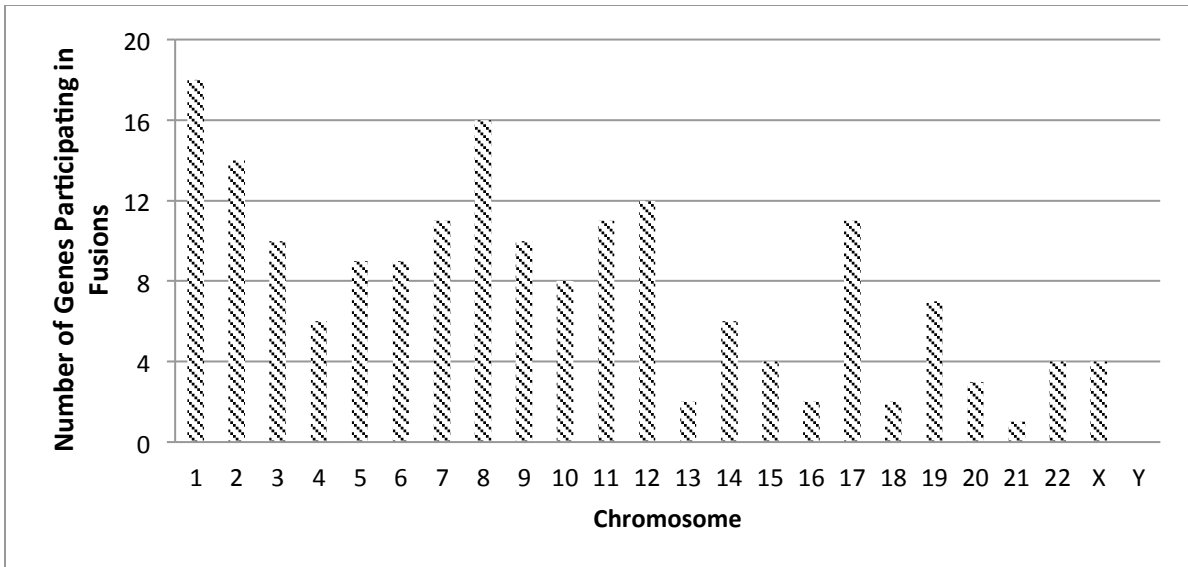


Figure 0-1: The distribution by chromosome position for the COSMIC reported genes participating in fusions.

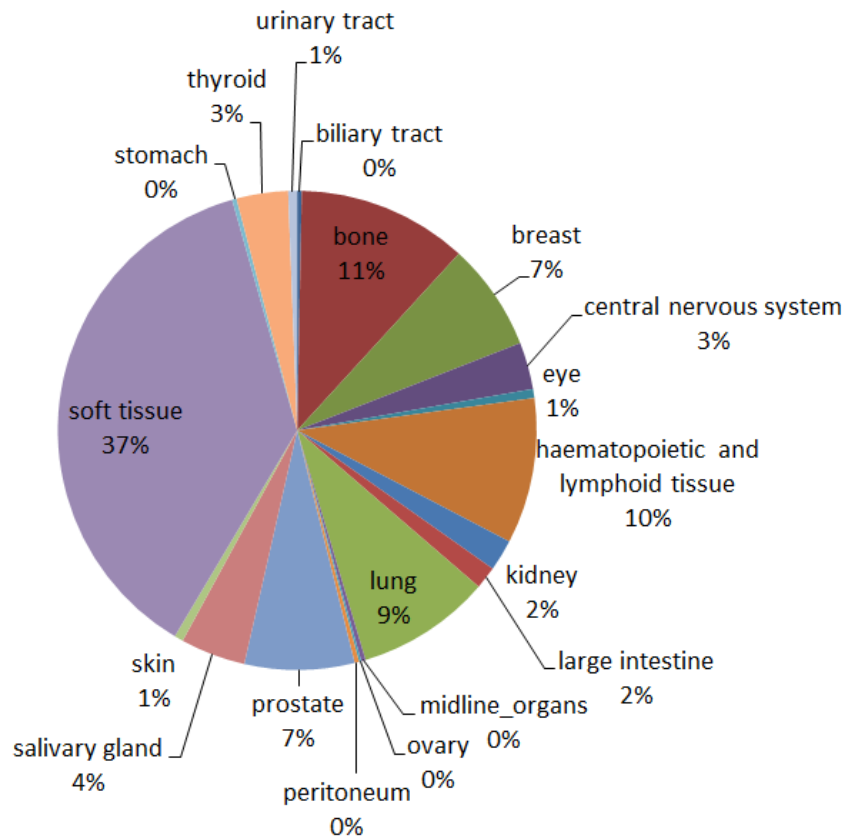


Figure 0-2: Percentages of COSMIC gene fusion records in given primary tissue types.

We filtered the gene fusion dataset for records with 5' and 3' gene names, mRNA transcripts, and first/last transcript nucleotide numbers. We excluded fusions involving more than two genes and those with additional insertion or deletion sequences. In Table 3, the number of unique sequences is higher than that of the unique gene pairs because, in some cases, a fusion pair exhibits different breakpoints. Sequence data is from the human genome reference 37.72, which we downloaded from Ensembl ([ftp://ftp.ensembl.org/pub/release-72/fasta/homo\\_sapiens/dna/](ftp://ftp.ensembl.org/pub/release-72/fasta/homo_sapiens/dna/)) (Flicek et al., 2014). The protein-protein interaction network data comes from PINA: Protein Interaction Network Analysis version 2 (Cowley et al., 2012).

## **2.2. Software and Tools**

All data processing was performed with programming languages Perl and R. Motif discovery was performed with Multiple EM for Motif Elicitation (MEME) version 4.9.0 (<http://ebi.edu.au/ftp/software/MEME/4.9.0/>) (Bailey et al., 2009). We chose this tool because it uses a deterministic optimization algorithm – it does not require a motif to exist in each input sequence and it determines likely motif lengths for the user (which is helpful when there are no a priori assumptions about potential motifs). We used TRANSFAC *F-Match* to search for statistically over-represented binding sites near breakpoints that form fusions (Matys et al., 2006). The network analysis was performed with the network construction and visualization software, Cytoscape (P. Shannon et al., 2003). Network statistics were calculated using the Cytoscape plug-in CentiScaPe (Scardoni, Petterlini, & Laudanna, 2009). The list of known protein-protein interactions was downloaded from PINA: Protein Interaction Network Analysis Platform version 2 (<http://cbg.garvan.unsw.edu.au/pina/download/>) (Cowley et al., 2012).

### 2.3. Gene fusion motif discovery

The goal for motif discovery in this study is to identify motifs enriched at gene fusion breakpoints, which help explain where they occur and why. The results may further help us predict novel fusion breakpoints in future. We performed *in silico* motif discovery on gene-fusion breakpoint sequences and compared results to motifs found at recombination hot spots and motifs identified by transcription factor binding sites.

To perform motif discovery on gene fusion breakpoints, it is necessary to know the nucleotide position on the genome where the break has occurred. Gene fusion data was collected from mRNA transcripts where introns were removed. If, upon mapping the mRNA transcripts back to the genome, we find that the breakpoint occurred within an exon, we know the exact position for the breakpoint. Otherwise, if we find that the breakpoint happened between two exons, we only know the intron region in which the breakpoint occurred. Since intron regions can be long, we did not find it appropriate to compare them to specific regions symmetric around an exonic breakpoint. Therefore, intron-region breakpoints were excluded from this study. It is more likely that gene fusions occur in intronic regions than in exonic regions – breakpoints in transcribed areas may be more disruptive and selected against. Gene-fusion discovery methods bias toward finding fusion with exonic breakpoints since they are easier to detect in mRNA transcripts. After filtering for exon-region breakpoints, we removed repeats and close-range breakpoints (those that exist more proximal to a breakpoint than the range of our desired excerpt length). Our input consisted of 142 genomic breakpoints that formed gene fusions.



Since some regulators act on regions several hundred base pairs away from where they bind, we strove to make our search space as wide as possible. One limiting factor was computational power. We found the maximum capacity for MEME with 142 input sequences is approximately 600 base pairs per sequence. Therefore, we extracted 600 base pair regions of genomic DNA centered on gene fusion breakpoints to serve as input. We want to uncover motifs that exist in a subset of gene fusions, which requires motif output beyond that of most statistical significance in the whole dataset. For this reason, we also extracted 300 base pair regions for input.

In our MEME command, we specified that sequences have zero or one occurrence of a given motif, that motifs may occur on the provided DNA strand or its reverse complement, and that the maximum motif length could be 20 base pairs (longer than the previously mentioned recombination hot spot motifs). MEME produced all summary statistics and graphs used in our analysis. We were most interested in motif E-values, which is an estimate for the expected number of motifs with its log likelihood ratio or higher that would exist in a like set of random sequences. Lower e-values denote greater significance. We also examined the position of the motifs within the sequences to determine position patterns (i.e. a given motif always exists upstream from breakpoint opposed to downstream).

Next, we chose to search for known transcription factor binding sites with the aim of uncovering enrichment at breakpoints that form fusions. Motif discovery should report enriched sites and obviate a targeted search, but the targeted search is beneficial in that it has much less probability for type II error. The targeted search is restricted to confirmed motifs

rather than motifs falling within a set or parameters. There is greater sensitivity for spotting those motifs, which is favorable in the event that motif discovery misses sequences with global maximum log likelihood. Of course, since the targeted search is restricted to known transcription factor binding sites, motif discovery remains beneficial when undocumented motifs exist. We used TRANSFAC's *F-Match* for the targeted search. Our input files included the breakpoint-centered 600 base pair sequences, plus 600 base pair control sequences centered at random exonic nucleotides within our gene set. We were interested in over represented sites with p-values less than 0.05.

#### **2.4. Gene fusion entropy analysis**

For the entropy analysis, we used 142 sequences of 600 base pairs centered at breakpoints that form gene fusions. These are the same sequences used in motif discovery – non-repeating, non-overlapping, exonic breakpoints. We also used 142 control sequences centered at random exonic breakpoints. We wrote a Perl program to set up the sliding frame and calculate entropies. We chose to take measurements in 20 base pair frames and shifted 5 base pairs with each measurement. Then, split by test and control, we averaged the reads in the same frame positions and plotted measurements on a line graph. We calculated local maximum and minimum reads and compared across the test and control set.

#### **2.5. Gene fusion network analysis**

We collected all gene fusion entries with 5' and 3' genes from the COSMIC. Unlike the motif discovery and entropy analysis, this study does not require sequence data so records without specific sequence data were included. Each unique gene fusion is represented once in

the network and there are 200 fusions pairs with 222 unique genes participating in fusions. We generated a control gene fusion dataset that complements our test set with 200 fusions and 222 unique genes by randomly sampling a list of all human genes.

We mapped the genes participating in fusions (in both the test and control data sets) to their respective proteins in a protein-protein interaction network and extracted all first-neighbor proteins. Then we built network files where each gene fusion is assigned a node and that node is connected to the union of proteins that interact with the 5' and 3' fusion genes. We used Cytoscape (P. Shannon et al., 2003) to visualize the networks and the Cytoscape plugin CentiScaPe (Scardoni et al., 2009) to calculate degree for all of the nodes. We compared the gene fusion degrees from the test network to those from the control network.

## CHAPTER III

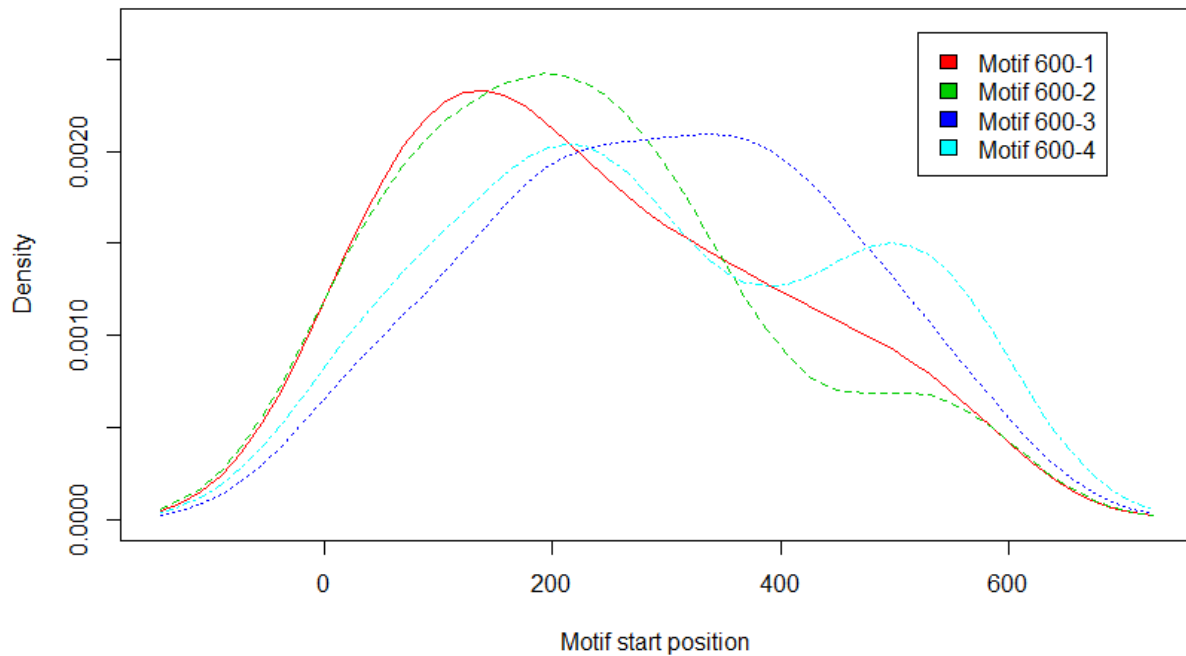
### Results

#### 3.1. Gene fusion motif discovery

The letter frequencies for the 600 base pair sequences input file is 0.283 A, 0.217 C, 0.217 G, and 0.283 T. These proportions served as the base frequencies for calculating significance for motifs in this dataset. Table 4 displays the motif output for the 600 base pair sequence file. Each of the four motifs are highly significant (e-value less than 0.05) but contain the maximum number of base pairs allotted. Motif number 600-4 has maximum conservation for G at nucleotides five and eight. Density maps of the motif sequence positions show homogenous distribution – there is no sharp skew upstream, downstream, or in a particular range for any of the four motif positions on the sequences (shown in Figure 0-1).

Table 4: Details on motifs found within 600 base-pair regions of breakpoints that form gene fusions.

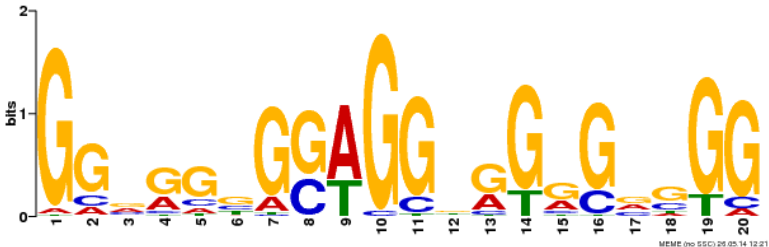
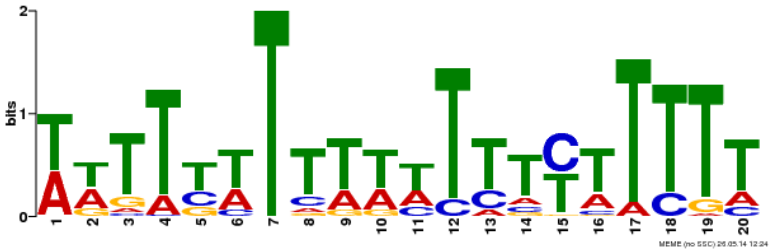
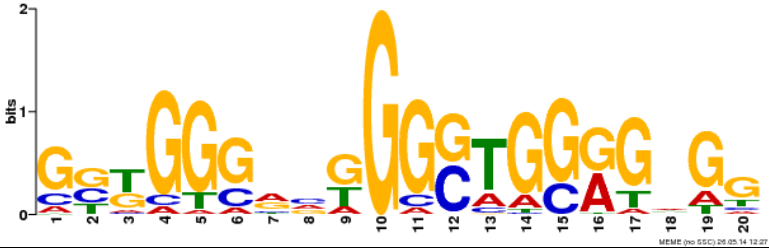
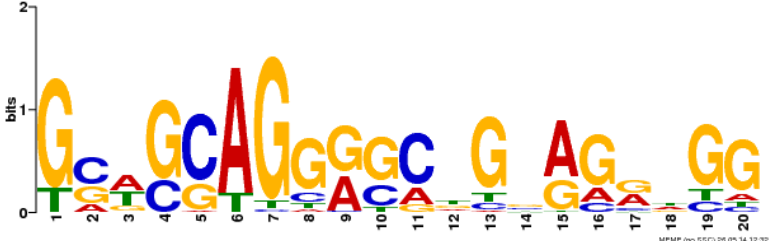
Motif	Percent of sequences exhibiting motif	E-value
<p><b>600-1.</b></p>	100%	1.9e-61
<p><b>600-2.</b></p>	57%	5.1e-61
<p><b>600-3.</b></p>	35%	3.4e-27
<p><b>600-4.</b></p>	36%	7.9e-12



**Figure 0-1:** Density plot for the start positions of motifs found in the 600 base pair breakpoint sequences. The x-axis represents nucleotide position and breakpoints occur in the center at 300 base pairs.

The letter frequencies for the 300 base pair sequences input file is 0.279 A, 0.221 C, 0.221 G, and 0.279 T. These proportions served as the base frequencies for calculating significance for motifs in this dataset. Table 5 displays the significant (e-value less than 0.05) motifs for the 300 base pair sequence file. Like results in the 600 base pair input file, all motifs are at the maximum 20 base pairs in length and the motif-sequence position maps show homogenous distribution for all motifs (shown in Figure 0-2). Motif number 600-1 and 300-2 are similar in that they are T-rich. However, other motifs from the two input files matched at high information content positions even though they have overlapping e-values.

Table 5: Details on motifs found within 300 base-pair regions of breakpoints that form gene fusions.

Motif	Percent of sequences exhibiting motif	E-value
<p><b>300-1.</b></p>  <p>bits</p> <p>MEME (no SSC) 25.05.14 12.21</p>	38%	2.4e-58
<p><b>300-2.</b></p>  <p>bits</p> <p>MEME (no SSC) 25.05.14 12.21</p>	44%	3.4e-45
<p><b>300-3.</b></p>  <p>bits</p> <p>MEME (no SSC) 25.05.14 12.21</p>	30%	2.9e-22
<p><b>300-4.</b></p>  <p>bits</p> <p>MEME (no SSC) 25.05.14 12.21</p>	26%	6.8e-5

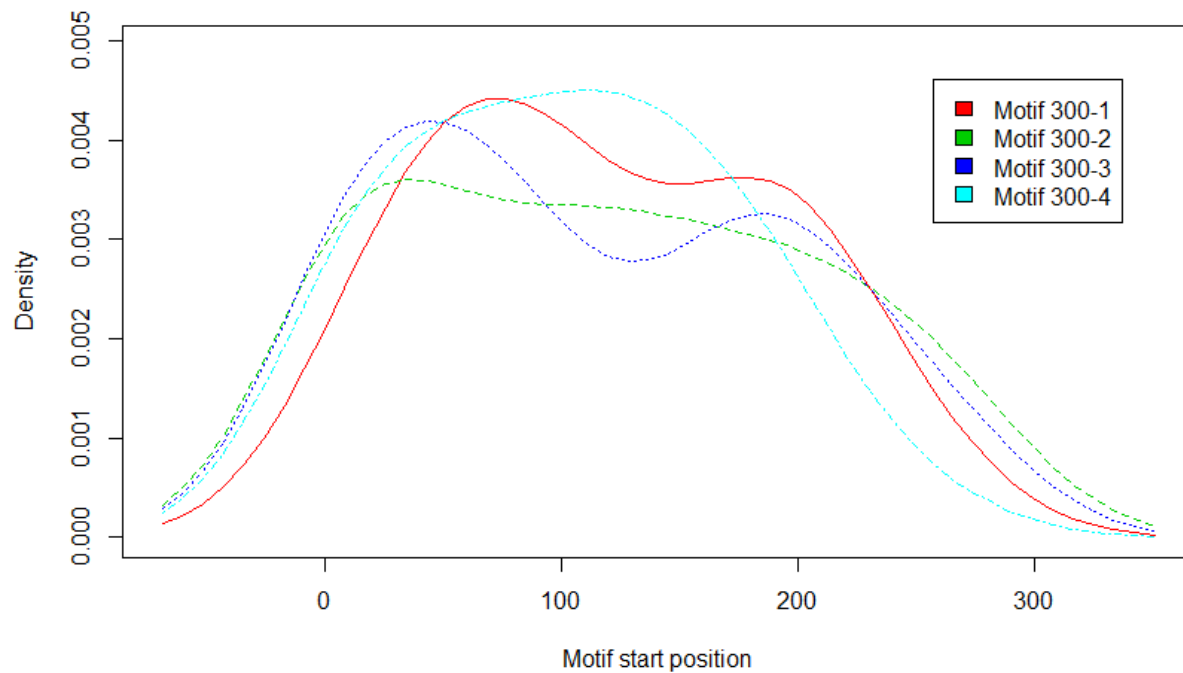


Figure 0-2: Density plot for the start positions of motifs found in the 300 base pair breakpoint sequences. The x-axis represents nucleotide position and breakpoints occur in the center at 150 base pairs.

The transcription factor binding site search by *F-Match* identified three sites over-represented in breakpoint regions that form gene fusions. Table 6 lists the sites with their test to control ratios and p-values. Each of the three sites had statistically significant p-values (less than 0.05).

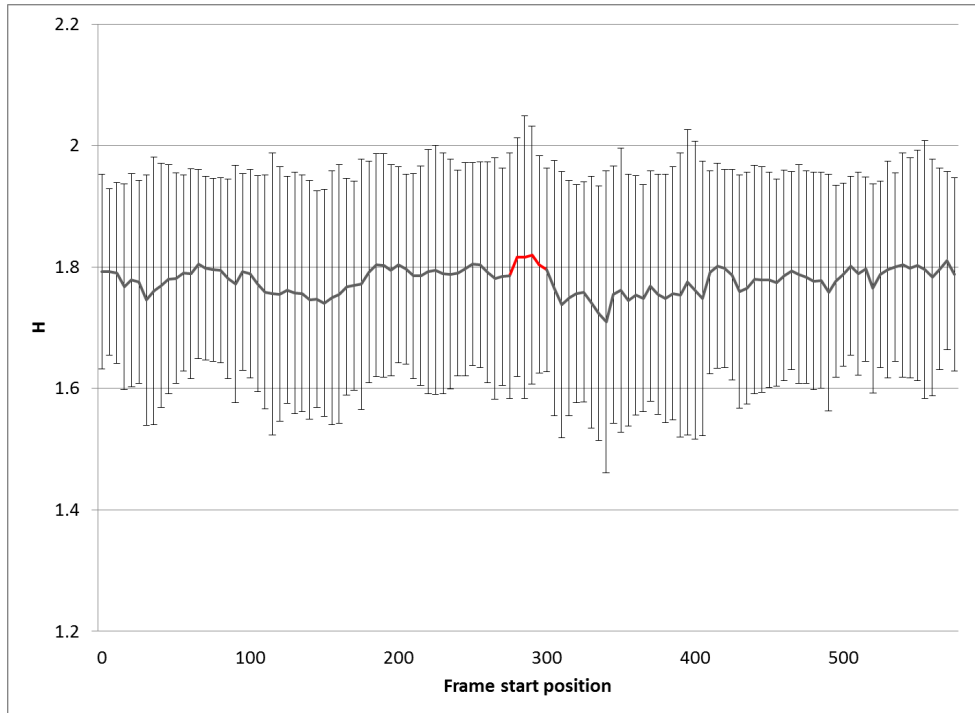
Table 6: Over-represented transcription factor binding sites in 600 base pair regions centered at breakpoints that form gene fusions compared to control sequences

Matrix Identifier	Name	Ratio (Test : Control)	P-value
V\$MINI19_B	Muscle initiator sequences-19	3.12	6.72e-4
V\$TATA_01	TATA Box	1.46	4.35e-4
V\$MUSCLE_INI_B	Muscle initiator	4.03	2.36e-4

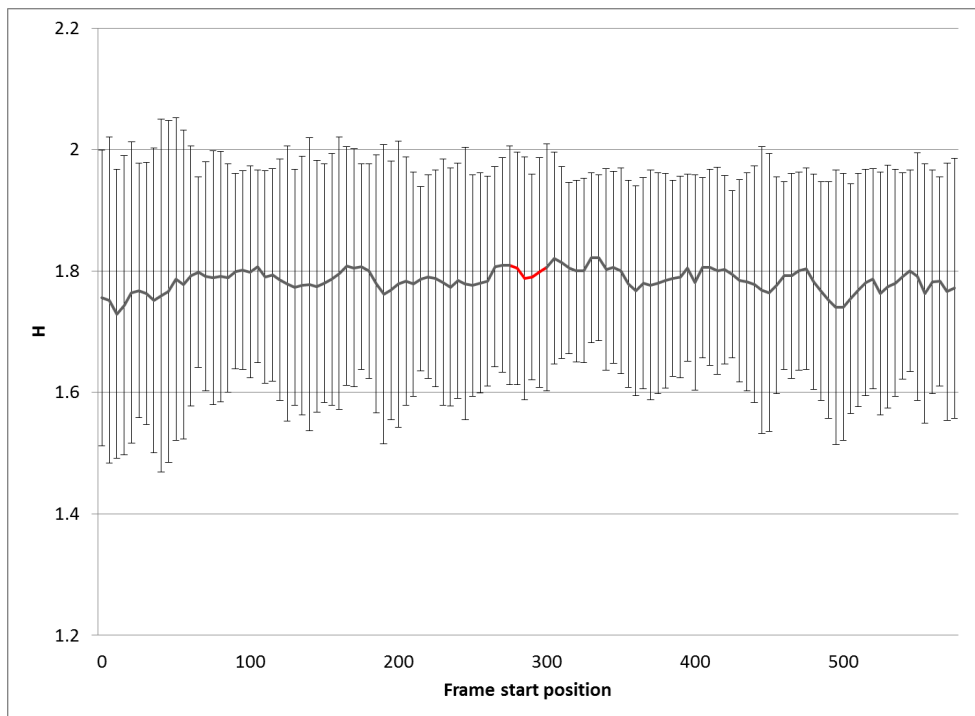


### **3.2. Gene fusion entropy analysis**

Figure 0-3 graphs the average entropy reads with standard error for 142 sequences centered at gene-fusion breakpoints. On average, the standard deviation is 10.4% of the entropy. The red portion of line represents reads the overlap breakpoints. The break occurs in a local high entropy area followed by a local low entropy area. Figure 0-4 graphs average entropy reads with standard error for 142 control sequences centered at random exonic breakpoints. On average, the standard deviation is 10.7% of the entropy. The red portion of line represents reads the overlap random breakpoints. Table 7 lists the maximum and minimum entropy reads for the test and control sets with their p-values from t-tests. The distance between the maximum and minimum average entropies is approximately 45 base pairs in the test set and 320 base pairs in the control set. The largest percent change in the test set is 17.29% over 45 base pairs and 15.15% over 45 base pairs in the control set.



**Figure 0-3: Average entropy values across 600 base pair sequences centered at gene fusion breakpoints. Reads are in 20 base pair frames with 5 bases per shift. The red fragment of line represents reads overlapping the breakpoints.**



**Figure 0-4: Average entropy values across 600 base pair control sequences centered at random exonic breakpoints. Reads are in 20 base pair frames with 5 bases per shift. The red fragment of line represents reads overlapping the random breakpoints.**

Table 7: The maximum and minimum average entropy reads for test and control sequences with p-values from t-tests.

	Test	Control	p-value
Max $H$	1.819	1.822	0.88
Min $H$	1.710	1.729	0.55
p-value	0.0004	0.0001	

### 3.3. Gene fusion network analysis

The gene fusion network is very large with 2,941 nodes and 14,812 edges. It is too complex to illustrate in a figure here. The degree distribution for the overall network follows the power-law (Figure 0-5) as does the degree distribution for only the nodes representing gene fusions (Figure 0-6). These results match those from Höglund et al. (Höglund et al., 2006) who also reported a scale-free gene fusion network.

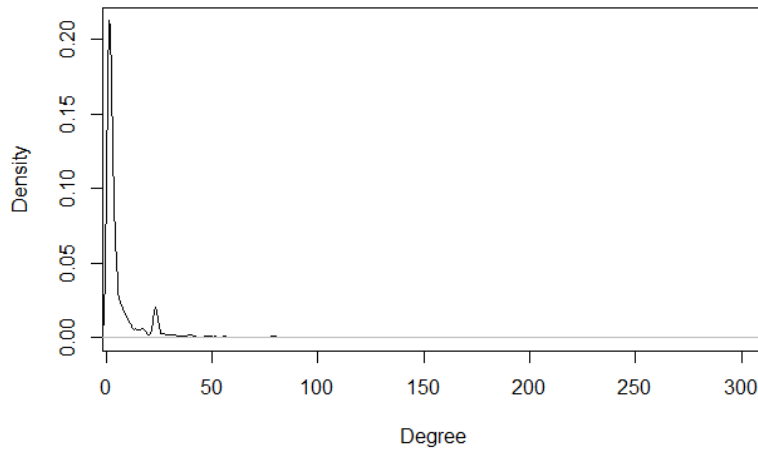
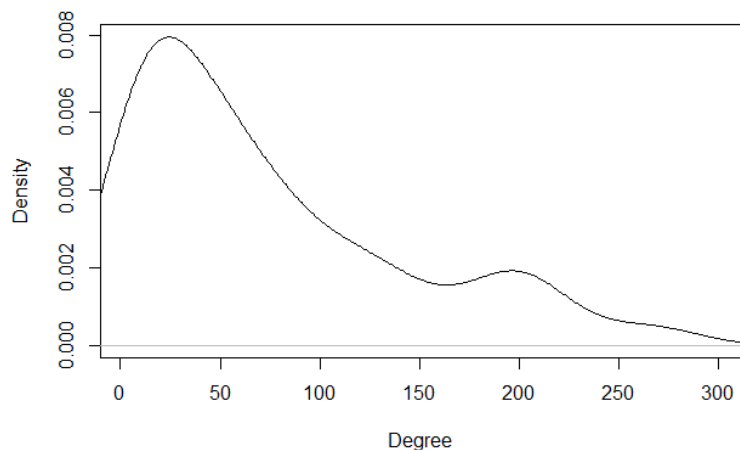
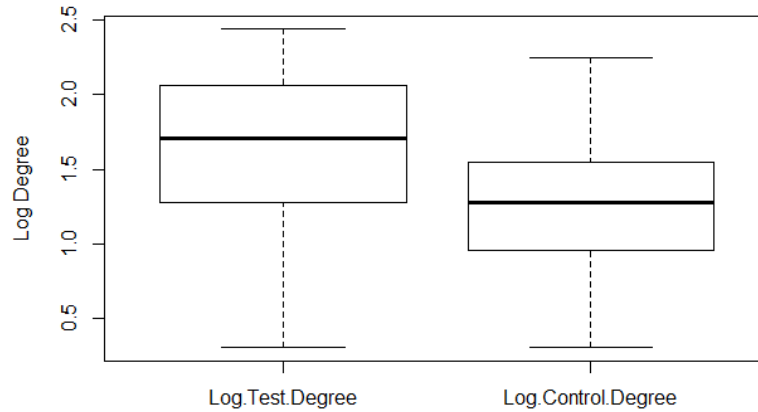


Figure 0-5: Degree distribution for the 2,941 nodes in the gene fusion network (fusions and protein interactors). The distribution follows a power-law, making the network scale-free.



**Figure 0-6: Degree distribution for the 200 gene fusion nodes in the network. The distribution follows a power-law.**

The control network has 2,593 nodes and 6,384 edges – 348 nodes and 8,428 edges less than the fusion network. This network is also scale-free. Since degree distribution follows a power law in these two networks, we log transformed the degrees to achieve normal distribution. Figure 0-7 shows a boxplot for the log transformed degree of nodes in the gene fusion test set compared to the control set. The test set average log degree is 1.635 with standard deviation 0.520 and the control set average log degree is 1.242 with standard deviation 0.440. The two averages are significantly different with a p-value less than 0.001 (calculated with a T-test).



**Figure 0-7: Box plot showing the degrees of nodes in the test gene fusion network and the control gene fusion network.**

RNA-binding protein EWS, coded by the Ewing sarcoma breakpoint region 1 (EWSR1) gene, has the most protein-protein interactions out of the set of resultant proteins in our gene fusion dataset. Gene fusions involving the gene EWSR1 have the highest degrees. Table 8 lists gene fusions with tenth percentile degrees in the network, which we define as hubs.

**Table 8: Gene fusions with tenth percentile degrees in the true gene fusion network.**

Rank	Gene Fusion	Degree	Rank	Gene Fusion	Degree
1	YY1/EWSR1	275	9	EWSR1/ATF1	207
1	EWSR1/YY1	275	10	EWSR1/PBX1	203
2	EWSR1/CREB1	272	11	EWSR1/WT1	203
3	PLAG1/CTNNB1	258	12	EWSR1/FLI1	203
3	CTNNB1/PLAG1	258	13	ACTB/GLI1	197
4	EWSR1/POU5F1	235	14	EWSR1/ETV1	197
5	EWSR1/SMARCA5	229	15	EWSR1/ERG	196
6	EWSR1/DDIT3	218	16	EWSR1/NFATC1	194
7	EWSR1/SP3	215	17	EWSR1/ZNF384	193
8	EWSR1/NFATC2	209	18	EWSR1/NR4A3	192
8	NFATC2/EWSR1	209	19	EWSR1/PATZ1	191
9	ATF1/EWSR1	207	20	EWSR1/ETV4	190

## CHAPTER IV

### Discussion

From these analyses we gained three overarching conclusions: 1) Motif discovery with broad parameters at mixed-cancer gene fusion breakpoints does not produce dependable output, 2) Gene fusion breakpoints likely occur at regions of relatively high entropy followed by regions of relatively low entropy, and 3) Network metrics may be useful toward understanding the role gene fusions have in cancers.

This work is novel by the following points. Previously, no large-scale motif discovery has been performed at breakpoints that form gene fusions, nor have there been sliding-window entropy analyses to determine patterns in entropy at breakpoints. While gene fusion networks have been built, our network is built from a large, composite, gene fusion dataset and our networks metrics are unique to that dataset.

#### **4.1. Gene fusion motif discovery**

Motif discovery at regions surrounding gene-fusion breakpoints yields eight statistically significant motifs from the 600 and 300 base pair datasets. Nevertheless, closer examination of the output shows evidence for overfitting, one of the limitations for motif discovery (Simcha, Price, & Geman, 2012). Each of the motifs contains the maximum number of base pairs permitted at run-time, which suggests generous substitutions were allowed to find a match. To address overfitting, Maclsaac et al. (Maclsaac & Fraenkel, 2006) suggested to perform motif

discovery on a fraction of the dataset followed by the evaluation of the motifs using the held-out data. This will yield an unbiased estimate of how well the motifs generalize to new data. MEME automatically performs this function, so as an alternative investigation for overfitting, we tried motif discovery against a similar size, similar GC content dataset of random sequences and looked for like significance values in any motifs found. Unfortunately, output from control sequences looked very similar to output from test sequences. We again saw highly significant 20-base-pair long motifs evenly distributed evenly across sequences. More evidence for overfitting comes from comparing the motifs and their e-values from the 600 base pair sequences and the 300 base pair sequences. None of the motifs across the two files matches although the e-values fall in the same range. If a motif from the 300 base pair sequence output was truly highly significant, it would certainly appear as a significant motif in the 600 base pair sequence output as well. Based on the results from random sequence motifs and the lack of motif overlap across 600 and 300 base pair test sequences, we conclude that motifs from our test set are not reliable.

It is likely that the size and range of the gene-fusion breakpoint datasets are in part responsible for poor performance from motif discovery. Nevertheless, there is currently no information that helps predict targeted regions where motifs exist in relation to gene-fusion breakpoints. Segregating the dataset into smaller ranges may reduce overfitting, but new complications arise. For example, if motifs exist in different ranges in different genes, they will be partitioned from one another, and not appear as statistically significant motifs in the split data set. Refining motif discovery for this dataset is an achievable task, but it requires planning to preempt potential drawbacks.

The three over-represented transcription factor binding sites near breakpoints are *muscle initiator sequences-19*, *TATA box*, and *muscle initiator*. The *TATA box* is ubiquitous and at a test to control ratio of 1.45, it is statistically enriched. *Muscle initiator sequences-19* and *muscle initiator* are two similar muscle promoters – they are both 21 bps long and 62% identical (13 matched nucleotides). This is an interesting finding that raises questions about gene fusion formation across tissue types. Does gene fusion formation occur by similar mechanisms regardless of tissue type or are there tissue-specific expedients? Although we know these two muscle-promoter sequences are enriched near breakpoints in this dataset, it is necessary to check for similar enrichment in a different dataset to make definitive conclusions. As COSMIC gathers additional gene fusion data and their cancer type of origin, we can test for the same enrichment pattern at new breakpoints, and thus gain support or dissuasion that these transcription factor binding sites are of interest in gene fusion research.

#### **4.2. Gene fusion entropy analysis**

The sliding entropy analysis shows a pattern that breakpoints occur at relatively high entropy regions followed by relatively low entropy regions. Nevertheless, the high entropy and low entropy regions are not significantly higher or lower than entropy measurements in control sequences. The difference between the test and control entropy reads is in the spatial distance between the maximum and minimum entropy sequences. In the test set, the change from the maximum to minimum entropy occurs over 45 base pairs, while in the control set the change occurs over 320 base pairs. The sharpest change in the control set is 15.15% over 45 base pairs, compared to 17.29% over 45 base pairs in the test set, and the percentages are significantly different with a p-value less than 0.0001.



Although error bars overlap, the difference between maximum and minimum entropy in the test set is significant with a p-value of 0.0004. The breakpoints occurring at relatively high entropy regions is in accordance with Gatlin's statement that low redundancy (high entropy) DNA molecules may have higher probability of error (Gatlin, 1968). A relatively high entropy region followed by a relatively low entropy region may be a structure that facilitates gene fusion formation.

Entropy signatures are less specific than motifs, but these results support that gene fusions do not occur randomly and are more likely to occur at region of relatively high entropy shortly followed by a region of relatively low entropy. This observation could be applied to gene fusion detection – gene fusion detection algorithms can include entropy-change checks to correct type II errors (false negatives). If a gene-fusion-detection algorithm based on sequence alone overlooks a gene fusion, a second pass checking changes in entropy could bring attention to the break for validation. However, before expanding gene-fusion-detection algorithms based on these findings, it is first necessary to collect gene fusion data from intronic breakpoints, and determine if the breaks happen at similar patterns. These results only define breaks in exonic regions, which is an unknown fraction of the total number of breakpoints that form fusions.

#### **4.3. Gene fusion network analysis**

The network analysis affirms two observations on gene fusion networks: gene fusion networks conform to power-law degree distributions found in naturally occurring networks, and average log degree is higher in true gene fusion networks compared to controls.

The gene-fusion network's degree distribution follows a power-law, therefore the network is scale-free rather than random (which would have a normal degree distribution) and contains hub genes. The hub genes in the network support that selection for gene fusion occurs during tumorigenesis. There are "promiscuous" genes that are more likely to form fusions in cancer, and they do so with a range of partners. If gene fusion formation is a random process, then there would be equal distribution for the genes involved in fusions. These results agree with the entropy analysis, which supports that gene fusion breakpoints do not occur at random sequences, but are more likely to occur at regions of relatively high entropy regions followed by relatively low entropy regions. If these regions are enriched in a specific gene, the gene will have a higher probability to form fusions.

Degree can be thought of as a measure for the impact a gene fusion has within the cell – the more interactions a gene's protein product has, the more disruptions will occur after fusion events. We reject the hypothesis that average log degrees are the same in observed-gene-fusion networks and networks built with randomly generated gene fusions with a p-value <0.001. We support the idea that gene fusions are more likely to be cancer-causal if the genes participating in the fusion interact with many proteins (higher than average degree).

Fourteen out of sixteen fusions with top ten highest degrees contain the gene EWSR1. EWSR1 encodes a protein involved in meiotic cell division, DNA repair mechanisms, and cellular ageing (Li et al., 2007). It is a promiscuous gene that form fusions with several partner genes in a variety of soft tissue tumors (reviewed in (Fisher, 2014)). The large number of EWSR1 fusions found in cancer tissue suggests driver ability and selection. The fusions with degrees falling in

the top ten that do not involve EWSR1 are those with pleiomorphic adenoma gene 1 (PLAG1) and catenin beta 1 (CTNNB1). PLAG1 encodes a zinc finger protein, and activation is a recurrent observation in lipoblastomas (Astrom et al., n.d.). CTNNB1 mutations are found in prostate cancer (Gerstein et al., 2002) and ovarian carcinomas (Palacios & Gamallo, 1998). High-degree gene fusions involve genes reported in literature with roles in cancer. The roles of the individual genes prior to fusion events are indicators of their role after gene fusion events. Therefore, the union of interactions for the resultant proteins in the gene fusion network acts as an indicator for the fusion's impact.

The gene fusion network is a useful tool for visualizing and measuring the impact a gene fusion has within the cell. The network shows that some genes (hubs) are more likely than others to form fusions, therefore could be positively selected for during tumorigenesis, and have more driver potential compared to fusions with less-common genes.

#### **4.4. Project Limitations**

The currently available algorithms for motif discovery have individual limitations. In an assessment on the DNA motif discovery algorithms, Simcha et al. reported that algorithms dependent on background models (like MEME) are “too null” which means they result in overly optimistic significance assessments (Simcha et al., 2012). This explains the overfitting that occurred in our analysis, and suggests that to reduce limitations, try alternative motif discovery tools and compare results.

The motif discovery and entropy analysis are limited by the reliability of the data sources. We depend on sequencing data and exact breakpoint positions in these studies, and

recognize that error in the reported nucleotide types and numbers can dampen results. COSMIC collects gene fusion records from various sources, and if the breakpoint positions are reported in different ways depending on the source (for example, the first nucleotide is number 0 from some sources and 1 from others), the entropy analysis loses some of its power. Motif results are less affected by small shifts because the sequences are aligned around the breakpoints. Nevertheless, incorrect sequencing data inhibits motif discovery algorithms from finding best fits, and may even lead to false positives if multiple incorrect sequences are incorporated. We very much depend on the data quality and integrity of others in the scientific community for these analyses.

Motif discovery, entropy analysis, and network analysis are all limited by our data set's size. COSMIC holds records for over nine thousand gene fusions, but less than 10% of those records have sequence information, and only a third of those are unique records. There is a plethora of unrecorded gene fusion information that, once collected, will give studies like these and many others a significant amount of power. As more information on gene fusions becomes publicly available, all of these studies can be repeated to yield results that are more dependable.

#### **4.5. Future Work**

We plan to improve upon the motif discovery analysis by repeating our steps with alternative motif discovery tools. In this study we utilized MEME suite, which uses a deterministic optimization algorithm. We found two alternative motif discovery tools with competitive performance: W-AlignACE, which uses Gibb sampling and a probabilistic

optimization algorithm (Chen, Guo, Fan, & Jiang, 2008), and Weeder, an enumerative motif discovery program (Pavesi, Mereghetti, Mauri, & Pesole, 2004). We hope that by comparing results from these three tools, we will gain power and be able to definitively conclude overfitting or lessen it. We plan to perform motif discovery at the amino acid level around exonic breakpoints. Degeneracy in the genetic code may lead to false negatives in motif discovery. Checking for motifs at the amino acid level may help reduce differences in the sequences caused by degeneracy.

We plan to expand out entropy analysis by validating our exonic-breakpoint-entropy-pattern against intronic-breakpoint-entropy-patterns. We must find data sources for gene fusions sequences from genomic DNA rather than mRNA to account for intronic breakpoints. Data is the limiting factor for this study, but we will collect sequence data as it becomes available. The rapidly evolving NGS technologies have helped investigators to generate more and more whole genome sequencing data, which will provide gene fusion breakpoints in intronic regions. If we are able to confirm our entropy pattern in both exonic and intronic breakpoints that form fusions, we can expand on a gene-fusion-detection algorithm by using entropy checks as a means to reduce false negatives in gene-fusion detection.

To further our network analysis, we plan to map Gene Ontology terms to genes participating in fusions and examine their functional enrichment. We believe this will highlight gene fusion groups common in cancer, like the tyrosine kinase fusions. Preliminary work for this step yielded hundreds of overrepresented Gene Ontology terms. Our current focus is reducing

noise by parsing the network into like-components (i.e. extracting subnetworks based on tissue of origin).

To advance this project, we plan to take advantage of the available gene-fusion sequence data and perform a survey on protein domain shuffling. We will explore the possibility of differentiating gene fusion drivers through shuffled substrate binding sites. We appreciate that the growing pool of gene fusion information incites creative investigations to answer questions about gene fusions and tumorigenesis.

## REFERENCES

- Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(Pt 21), 4947–57. doi:10.1242/jcs.02714
- Aman, P. (1999). Fusion genes in solid tumors. *Seminars in Cancer Biology*, 9(4), 303–18. doi:10.1006/scbi.1999.0130
- Astrom, A., D'Amore, E. S., Sainati, L., Panarello, C., Morerio, C., Mark, J., & Stenman, G. (n.d.). Evidence of involvement of the PLAG1 gene in lipoblastomas. Spandidos Publications. Retrieved from <http://www.ingentaconnect.com/content/sp/ijo/2000/00000016/00000006/art00004>
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue), W202–8. doi:10.1093/nar/gkp335
- Barilá, D., & Superti-Furga, G. (1998). An intramolecular SH3-domain interaction regulates c-Abl activity. *Nature Genetics*, 18(3), 280–2. doi:10.1038/ng0398-280
- Ben-Neriah, Y., Daley, G., Mes-Masson, A., Witte, O., & Baltimore, D. (1986). The chronic myelogenous leukemia-specific P210 protein is the product of the bcr/abl hybrid gene. *Science*, 233(4760), 212–214. doi:10.1126/science.3460176
- Blekas, K., Fotiadis, D. I., & Likas, A. (2003). Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics (Oxford, England)*, 19(5), 607–17. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12651719>
- Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., da Piedade, I., ... Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research*, 36(Database issue), D102–6. doi:10.1093/nar/gkm955
- Chen, X., Guo, L., Fan, Z., & Jiang, T. (2008). W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics (Oxford, England)*, 24(9), 1121–8. doi:10.1093/bioinformatics/btn088
- Chmielecki, J., Peifer, M., Jia, P., Socci, N. D., Hutchinson, K., Viale, A., ... Pao, W. (2010). Targeted next-generation sequencing of DNA regions proximal to a conserved GXGXXG signaling motif enables systematic discovery of tyrosine kinase fusions in cancer. *Nucleic Acids Research*, 38(20), 6985–96. doi:10.1093/nar/gkq579
- Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., ... Wu, J. (2012). PINA v2.0: mining interactome modules. *Nucleic Acids Research*, 40(Database issue), D862–5. doi:10.1093/nar/gkr967

- D'haeseleer, P. (2006). What are DNA sequence motifs? *Nature Biotechnology*, 24(4), 423–5. doi:10.1038/nbt0406-423
- Dempster, A. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series A. General*, 39(1).
- Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A., & Ziv, J. (1995). On the entropy of DNA: algorithms and measurements based on memory and rapid convergence, 48–57. Retrieved from <http://dl.acm.org/citation.cfm?id=313651.313662>
- Fisher, C. (2014). The diversity of soft tissue tumours with EWSR1 gene rearrangements: a review. *Histopathology*, 64(1), 134–50. doi:10.1111/his.12269
- Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., ... Searle, S. M. J. (2014). Ensembl 2014. *Nucleic Acids Research*, 42(Database issue), D749–55. doi:10.1093/nar/gkt1196
- Forbes, S. A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., ... Futreal, P. A. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Research*, 38(Database issue), D652–7. doi:10.1093/nar/gkp995
- Freeman, L. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1), 35 – 41.
- Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., ... Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews. Cancer*, 4(3), 177–83. doi:10.1038/nrc1299
- Gatlin, L. L. (1968). The information content of DNA. II. *Journal of Theoretical Biology*, 18(2), 181–194. doi:10.1016/0022-5193(68)90160-4
- Ge, H., Liu, K., Juan, T., Fang, F., Newman, M., & Hoeck, W. (2011). FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics (Oxford, England)*, 27(14), 1922–8. doi:10.1093/bioinformatics/btr310
- Gerstein, A. V, Almeida, T. A., Zhao, G., Chess, E., Shih, I.-M., Buhler, K., ... Papadopoulos, N. (2002). APC/CTNNB1 (beta-catenin) pathway alterations in human prostate cancers. *Genes, Chromosomes & Cancer*, 34(1), 9–16. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11921277>
- Höglund, M., Frigyesi, a, & Mitelman, F. (2006). A gene fusion network in human neoplasia. *Oncogene*, 25(18), 2674–8. doi:10.1038/sj.onc.1209290
- Kim, P., Yoon, S., Kim, N., Lee, S., Ko, M., Lee, H., ... Lee, S. (2010). ChimerDB 2.0--a knowledgebase for fusion genes updated. *Nucleic Acids Research*, 38(Database issue), D81–5. doi:10.1093/nar/gkp982
- Koslicki, D. (2011). Topological entropy of DNA sequences. *Bioinformatics (Oxford, England)*, 27(8), 1061–7. doi:10.1093/bioinformatics/btr077
- Kumar-Sinha, C., Tomlins, S. A., & Chinnaiyan, A. M. (2008). Recurrent gene fusions in prostate cancer. *Nature Reviews. Cancer*, 8(7), 497–511. doi:10.1038/nrc2402



- Li, H., Watford, W., Li, C., Parmelee, A., Bryant, M. A., Deng, C., ... Lee, S. B. (2007). Ewing sarcoma gene EWS is essential for meiosis and B lymphocyte development. *The Journal of Clinical Investigation*, *117*(5), 1314–23. doi:10.1172/JCI31222
- Liu, C., Ma, J., Chang, C. J., & Zhou, X. (2013). FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq. *BMC Bioinformatics*, *14*, 193. doi:10.1186/1471-2105-14-193
- Maclsaac, K. D., & Fraenkel, E. (2006). Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Computational Biology*, *2*(4), e36. doi:10.1371/journal.pcbi.0020036
- Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., ... Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, *34*(Database issue), D108–10. doi:10.1093/nar/gkj143
- McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M. G. F., ... Shah, S. P. (2011). deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Computational Biology*, *7*(5), e1001138. doi:10.1371/journal.pcbi.1001138
- Mitelman, F., Johansson, B., & Mertens, F. (2004). Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature Genetics*, *36*(4), 331–4. doi:10.1038/ng1335
- Mizuno, T., Kyoizumi, S., Suzuki, T., Iwamoto, K. S., & Seyama, T. (1997). Continued expression of a tissue specific activated oncogene in the early steps of radiation-induced human thyroid carcinogenesis. *Oncogene*, *15*(12), 1455–60. doi:10.1038/sj.onc.1201313
- Myers, S., Freeman, C., Auton, A., Donnelly, P., & McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, *40*(9), 1124–9. doi:10.1038/ng.213
- Nambiar, M., Kari, V., & Raghavan, S. C. (2008). Chromosomal translocations in cancer. *Biochimica et Biophysica Acta*, *1786*(2), 139–52. doi:10.1016/j.bbcan.2008.07.005
- Nikiforov, Y. E., & Nikiforova, M. N. (2011). Molecular genetics and diagnosis of thyroid cancer. *Nature Reviews. Endocrinology*, *7*(10), 569–80. doi:10.1038/nrendo.2011.142
- Novo, F. J., de Mendíbil, I. O., & Vizmanos, J. L. (2007). TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics*, *8*, 33. doi:10.1186/1471-2164-8-33
- O'Brien, S. G., Guilhot, F., Larson, R. A., Gathmann, I., Baccarani, M., Cervantes, F., ... Druker, B. J. (2003). Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *The New England Journal of Medicine*, *348*(11), 994–1004. doi:10.1056/NEJMoa022457
- P. C. Nowell, D. A. H. (1960). A Minute Chromosome in Human Chronic Granulocytic Leukemia. *Science*.

- Palacios, J., & Gamallo, C. (1998). Mutations in the {beta}-Catenin Gene (CTNNB1) in Endometrioid Ovarian Carcinomas. *Cancer Res.*, 58(7), 1344–1347. Retrieved from <http://cancerres.aacrjournals.org/content/58/7/1344.short>
- Pavesi, G., Mereghetti, P., Mauri, G., & Pesole, G. (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research*, 32(Web Server issue), W199–203. doi:10.1093/nar/gkh465
- Pingoud, A., & Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. *Nucleic Acids Research*, 29(18), 3705–27. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=55916&tool=pmcentrez&rendertype=abstract>
- Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4), 581–603. doi:10.1007/BF02289527
- Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., ... Gerstein, M. B. (2010). FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biology*, 11(10), R104. doi:10.1186/gb-2010-11-10-r104
- Scardoni, G., Petterlini, M., & Laudanna, C. (2009). Analyzing biological network parameters with CentiScaPe. *Bioinformatics (Oxford, England)*, 25(21), 2857–9. doi:10.1093/bioinformatics/btp517
- Schneider, T. D., Stormo, G. D., Gold, L., & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3), 415–31. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3525846>
- Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1), 3. doi:10.1145/584091.584093
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–504. doi:10.1101/gr.1239303
- Shaw, A. T., Hsu, P. P., Awad, M. M., & Engelman, J. A. (2013). Tyrosine kinase gene rearrangements in epithelial malignancies. *Nature Reviews. Cancer*, 13(11), 772–87. doi:10.1038/nrc3612
- Shugay, M., Ortiz de Mendíbil, I., Vizmanos, J. L., & Novo, F. J. (2013). Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics (Oxford, England)*, 29(20), 2539–46. doi:10.1093/bioinformatics/btt445
- Simcha, D., Price, N. D., & Geman, D. (2012). The limits of de novo DNA motif discovery. *PLoS One*, 7(11), e47836. doi:10.1371/journal.pone.0047836
- Sinha, S., & Tompa, M. (2002). Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 30(24), 5549–60. Retrieved from

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=140044&tool=pmcentrez&rendertype=abstract>

- Thieme, S., & Groth, P. (2013). Genome Fusion Detection: a novel method to detect fusion genes from SNP-array data. *Bioinformatics (Oxford, England)*, 29(6), 671–7. doi:10.1093/bioinformatics/btt028
- Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., & Moreau, Y. (2002). A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, 9(2), 447–64. doi:10.1089/10665270252935566
- Wang, X.-S., Prensner, J. R., Chen, G., Cao, Q., Han, B., Dhanasekaran, S. M., ... Chinnaiyan, A. M. (2009). An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nature Biotechnology*, 27(11), 1005–11. doi:10.1038/nbt.1584
- Wong, S., & Witte, O. N. (2004). The BCR-ABL story: bench to bedside and back. *Annual Review of Immunology*, 22, 247–306. doi:10.1146/annurev.immunol.22.012703.104753
- Wu, C.-C., Kannan, K., Lin, S., Yen, L., & Milosavljevic, A. (2013). Identification of cancer fusion drivers using network fusion centrality. *Bioinformatics (Oxford, England)*, 29(9), 1174–81. doi:10.1093/bioinformatics/btt131

## APPENDIX

### Code for sliding-window entropy calculations

```
#!/usr/bin/perl
#
#This program designates a window of width $FL (frame length) in which to
calculate entropy
#and move the window along input sequences shifting width $SL (shift length)
along the desired
#region 2*DL (desired length) centered at a breakpoint
#
#Author: Morgan Harrell

#FRAME LENGTH
$FL = 20;
#SHIFT LENGTH
$SL = 5;
#HALF LENGTH DESIRED AROUND BREAKPOINT
$DL = 300;

#Open exon breakpoint ID list
#Form:
#mRNA,breakpoint position, gene name_chromosome
#NM_004083.4,837,DDIT3_12
open (EID, "exonBP.info.NoRepeats") || die ("Cannot find input list\n");
while (<EID>)
{
    chomp($_);
    my @values = split(' ', $_);
    #Get DNA sequence for the gene with an exonic breakpoint
    open (mRNA, "/gpfs22/home/harrelm/data/COSMIC/geneSeq/genes/$values[2]")
|| print "Cannot find
/gpfs22/home/harrelm/data/COSMIC/geneSeq/genes/$values[2]\n";
    my @sequence = <DNA>;
    close(DNA);

    #Prep output
    print "$values[2]\t";

    #Set up slider loop
    my $length = length($sequence[1]);
    my $start = $values[1] - $DL;
    #Prevents negative number in the event that the breakpoint is not more
than DL nucleotides
    #into the sequence
    if ($start < 0)
    {
        $start = 0;
    }
    my $end = $values[1] + $DL;
```

```

    #Prevents overarching number in the event that the breakpoint is not more
    than DL nucleotides
    #from the end of the sequence
    if ($start > $length)
    {
        $start = $length;
    }

    #Slide and calculate entropy
    for (my $i = $start; $i+$FL < $end; $i+=$SL)
    {
        my $seqSub = substr($sequence[1], $i, $FL);
        #Sequence entropy
        my $sse = &entropy($seqSub);
        #Print entropy
        print "$sse\t";
    }
    print "\n";
}
close(EMID);

sub entropy
{
    my $string = $_[0];
    my %Count;
    my $total = 0;
    foreach my $char (split(/./, $string))
    {
        $Count{$char}++;
        $total++;
    }
    my $H = 0;
    foreach my $char (keys %Count)
    {
        my $p = $Count{$char}/$total;
        $H += $p * log($p);
    }
    $H = -$H/log(2);
}

```