A Systems Biology Approach to Characterizing Gene Fusion Pathways in Cancer

Tressa R. Hood

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2017

Committee:

Neil Abernethy

Erin Piazza

Ali Shojaie

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

A Systems Biology Approach to Characterizing Gene Fusion Pathways in Cancer

Tressa R. Hood

Chair of the Supervisory Committee:
Neil Abernethy, PhD, Associate Professor
Biomedical and Health Informatics

Gene fusions have long been known to drive cancer. Initial discovery of gene fusions was opportunistic, and functional assessment was done individually and experimentally. There is no comprehensive systems biology approach to understanding the impact of gene fusions on the signaling networks within tumor cells. An integrative computational approach was taken to achieve a better understanding of gene fusions and their complex influence on pathways and interaction networks in the context of lung cancer. Using well-studied fusions and publicly available gene expression data, the effect of fusion events on the expression pattern of gene networks revealed unique differences in tumors with gene fusions, tumors without gene fusions, and normal samples. This approach identifies gene expression signatures associated with specific fusions, and provides a model for integrating experimental and pathway data to better understand the biology of a fusion genes and their roles in oncogenesis.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like express my heartfelt gratitude to the members of my committee, Drs. Neil Abernethy, Erin Piazza, and Ali Shojaie, for their invaluable contributions to this work. I would also like to thank Lucy L. Wang, Aakash Sur and other students for their patience, insights and general willingness to listen to my thoughts on this project at various times.

INTRODUCTION

The "hallmarks of cancer" were proposed over a decade ago and have provided an invaluable conceptual framework for the study of oncology[1]. These hallmarks represent crucial modifications in cell function that result in malignancy: sustaining proliferative signaling, deregulating cellular energetics, resisting cell death, and activating invasion and metastasis, among others[1]. Underlying many of these attributes is genomic instability, the increased rate of mutations during the cell cycle and the cause of chromosomal aberrations known as fusion genes[2]. Gene fusions occur when two genes merge to form a hybrid gene either by translocations, interstitial deletions, or chromosomal inversions [Fig. 1]. Fusions can lead to tumorigenesis by activating cancer-causing proto-oncogenes either through translocating them downstream of a strong promotor or altering its protein structure[3]. In recent years, gene fusions have been increasingly detected among common solid tumors, and have been found to play an important role in lung adenocarcinoma[4]. These discoveries underscore the clinical importance of fusion genes, and suggest their potential use in diagnosis, prognosis, and personalization of cancer treatment[5].

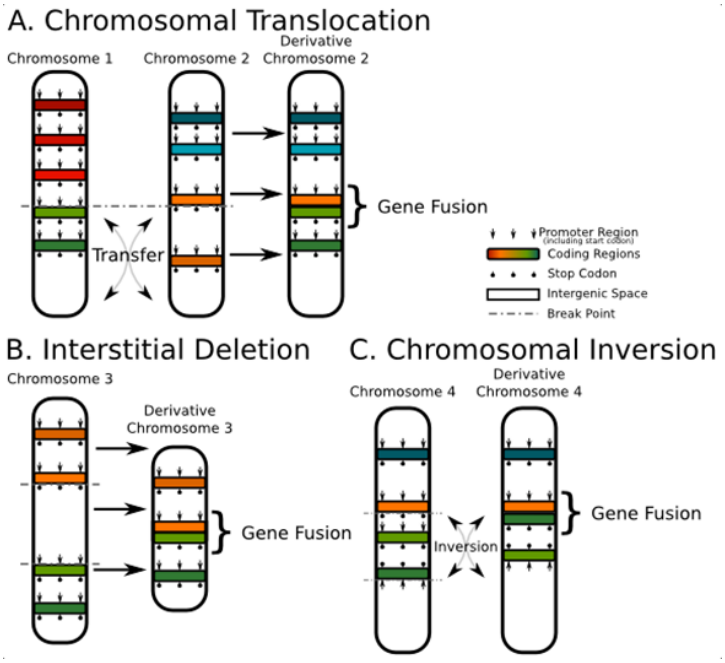


Figure 1. Gene Fusion Formation (A) Chromosomal translocation, which is the rearrangement of parts between non-homologous chromosomes. (B) Interstitial deletions, are deletions that do not involve the terminal end of the chromosome. (C) Chromosomal inversions, where the chromosome is reversed end to end[47].

Gene Fusions

Gene fusion research started with the discovery of the "Philadelphia Chromosome" in the early 1980s, which is a fusion of the BCR and ABL1 genes, that results from a translocation between chromosomes 9 and 22[4]. This fusion presents in 90% of patients with chronic myeloid leukemia (CML)[4]. Initially, identification of gene fusions was performed via fluorescence *in situ* hybridization (FISH) and real-time polymerase chain reaction (RT-PCR). More recently, next-generation sequencing and DNA microarrays can detect fusions in a high-throughput manner. Most studies involving fusions focus on detection using RNA-seq or whole genome sequencing data (WGS)[6,7,8]. In contrast, few studies have gone beyond detection, and explored the cellular impacts of gene fusions. Latysheva et al., merged molecular characterization, identification, and clinical significance of gene fusions, and found that fusions inhabit central positions in the interaction networks of clinically relevant genes[8,9]. Their approach shows that we can leverage existing knowledge resources to holistically understand the overarching biological principles that exist between fusions.

Molecular pathways

Pathways and interaction networks are an important framework for understanding the complexities of cancer[8,9,10]. A pathway is a series of interactions among different molecules within the cell that leads to a modification of the cell or the creation of a product[10]. The most common pathways studied are involved in the regulation of gene expression, metabolism, and signal transduction[10]. Often, abnormalities can wreak havoc on tightly regulated pathways such as apoptosis and replication, key players in cancer[8,9]. The regulation of these pathways is critical to keep healthy cells alive and eliminate aberrant cells[11].

Gene Interaction Networks

A gene interaction network is a representation of the dyadic interactions among a group of genes, where each dyad sums the physical and functional relationships between a pair of genes[12]. The impacts of gene fusions may not be limited to canonical pathways, and may have a more subtle influence on the larger gene interaction network[13]. The more we understand disease-state gene interactions, the more we can understand the role of novel gene fusions and predict their function.

The impact of gene fusions on gene network interactions has been understudied[8,9]. Thus far, progress has been hindered by lack of clear functional annotations of gene partners involved in fusions. Typically, one of the fusion partners is an oncogene, often backed by a wealth of knowledge, while the other is an obscure gene only of note due to its participation in the fusion. However, Wu et al. leverages knowledge of gene interactions to determine gene fusions that are more likely to drive cancer[14]. They proposed a method which estimates whether a novel fusion will be an oncogenic driver based on the location of the partner gene in an interaction network; notably, hubs are more likely to drive tumorigenesis than genes at any other position in a network[14]. By leveraging the gene interaction networks and annotations of the surrounding neighbor genes, it can be possible to infer the functionality and effects of the unknown partner genes[14].

GAPS IN GENE FUSION RESEARCH

While the primary focus of research has been on detection of known fusions from sequencing data and identifying novel fusions, understanding the biology of these genomic abnormalities is critical. Current strategies do not leverage the vast amounts of information spawned by the genomic revolution. Currently there is no systems biology approach to identify the impact of gene fusions on cell physiology, regulatory pathways, or interaction networks.

In addition, there is a lack of a comprehensive approach that leverages and combines the many existing informatic tools. There are extensive pathway databases including the Kyoto Encyclopedia of Genes and Genomes Pathway Database (KEGG Pathway)[15,16], Reactome[17], and Gene Ontology (GO)[18], and interaction databases, most notably bioGRID[13]. Furthermore, there are interactive tools available like Cytoscape[19], that allow you to integrate, analyze, and visualize gene network data. Finally, there are software packages and algorithms available to detect gene fusions from RNA-seq data, e.g. deFuse[7], INTEGRATE[6], and TopHat-Fusion[20]. While there are many tools that excel at individual tasks, to our knowledge, there is nothing that synthesizes them.

Finally, pathway analysis in the presence of gene fusions requires special considerations because the topology of the network changes as a result of the fusion event[8,9]. Functions can be gained or lost because of the physical joining of two genes and often concomitant loss of functional domains, which impacts both direct and indirect interactions of these proteins with their networks. Gene fusions can also impact direct interactions, whether it be by novel interactions or a change in the regulation of an interaction.

Characterizing gene fusions and investigating their impacts on pathways and gene interactions can lead to a greater understanding of the malignancies driven by these fusions as well

as open new avenues for treatment. Without a workflow to integrate all this information, gene fusions and the effects of their influence will continue to be studied individually and the connections between different fusions and their corresponding diseases will elude us.

RESEARH AIMS

Genomic instability has been known to be a major driving force in tumorigenesis[21]. The transformation of healthy cells to cancerous ones can vary across different types of cancer, however, there is a convergence on modifications that occur on basic cellular functions (proliferation and apoptosis)[1,21]. Gene fusions resulting in alterations to proliferation and cell death pathways should be detectable not only at the sequence and gene expression level but also at the pathway and interaction network level. Therefore, we hypothesize that computational methods can characterize the impact that gene fusions have on pathways.

To evaluate this hypothesis, we developed a comprehensive computational workflow to investigate gene fusions. Our approach evaluates an individual gene fusion by identifying the differentially expressed genes it causes, defining the regulatory pathways that are unique to these genes, categorizing their local interaction network, and showing the gene expression changes between those interactions [Fig. 2]. We believe our results may shed light on how gene fusions effect the regulatory pathways and gene interaction networks that lead to cancer.

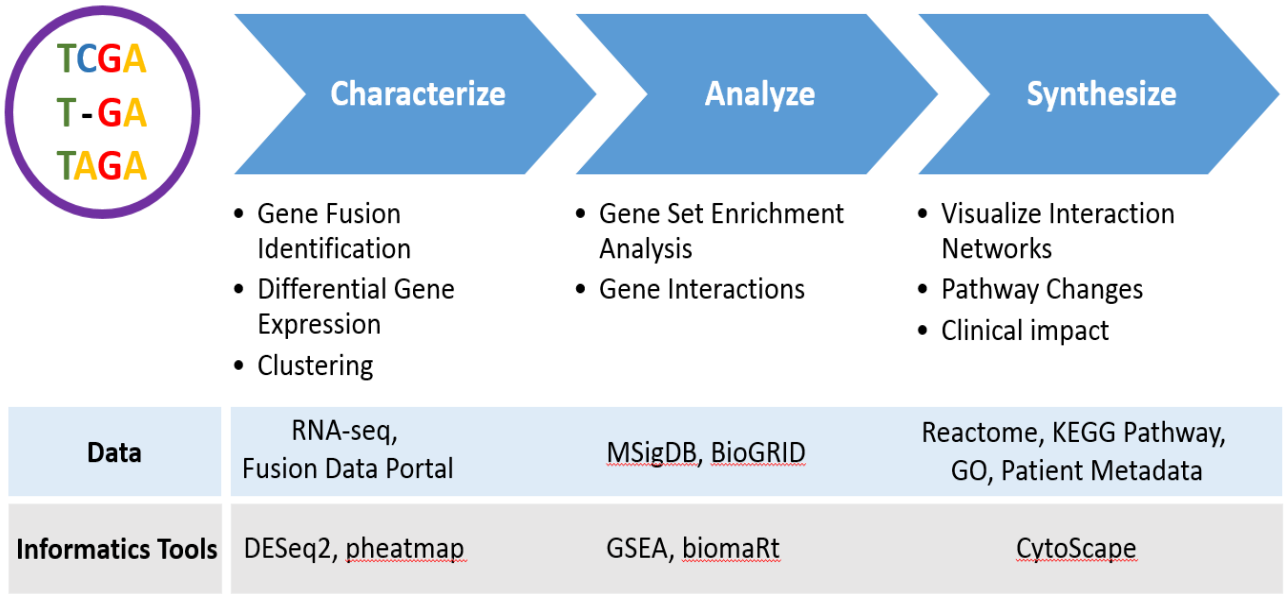


Figure 2. Gene Fusion Analysis Workflow. We developed a comprehensive computational workflow to investigate gene fusions. Our approach involves three comprehensive steps: characterize, analyze and synthesize. In the characterize step, we use multiple informatics tools to identify differentially expressed genes from RNA-seq data, and cluster by different phenotypes. We then investigate these genes to determine the pathways and interactions they are involved with by using Gene Set Enrichment Analysis (GSEA)[29]. Finally, we synthesize the data by visualizing the interaction networks, evaluating for pathway changes due to the fusion event and assessing the clinical impacts.

To determine the validity of the approach proposed, we chose to focus on one particular fusion, EML4-ALK in lung adenocarcinoma (LUAD). It is both clinically impactful and biologically relevant and is among the most established fusions in lung cancer as it presents in approximately 4-6% of LUAD patients [22]. In addition, The Cancer Genome Atlas (TCGA)[23] has a well annotated and comprehensive LUAD RNA-seq dataset.

To investigate this approach and hypothesis, we propose the following research questions:

1. Are fusion samples distinct from normal tissue and fusion-free tumors?
2. How do gene fusions change the expression of pathways?
3. How do fusions change the interaction of pathways?

The goal of this work is to address these questions using the EML4-ALK fusion in lung cancer as a proof-of-principle for the application of a computational approach for analyzing the impact of gene fusions on functional networks in cancer signaling.

METHODS

The Cancer Genome Atlas & Fusion Detection

The Cancer Genome Atlas (TCGA) was used as a starting point to gather RNA-seq datasets for lung adenocarcinoma (LUAD)[23]. TCGA is a large open source database of de-identified clinical and biological data for over 11,000 patients[23]. The LUAD subset consisted of 594 patients, including 59 patients matched normal tissue samples.

We also explored different fusion-calling algorithms to determine presence of fusions in this RNA-seq dataset. After reviewing several well-known algorithms, aided by the comprehensive evaluation of 12 different fusion detection software packages provided by Kumar et al, the Pipeline for RNA sequencing Data Analysis (PRADA) was selected for its high sensitivity, specificity, and computational efficiency[24,25,26]. The algorithm detects fusions by identifying discordant read pairs and evaluating fusion junction spanning reads [Fig. 3][25,26]. We utilized this algorithm and the accompanying public dataset of known and novel fusions the authors identified in TCGA[25,26].

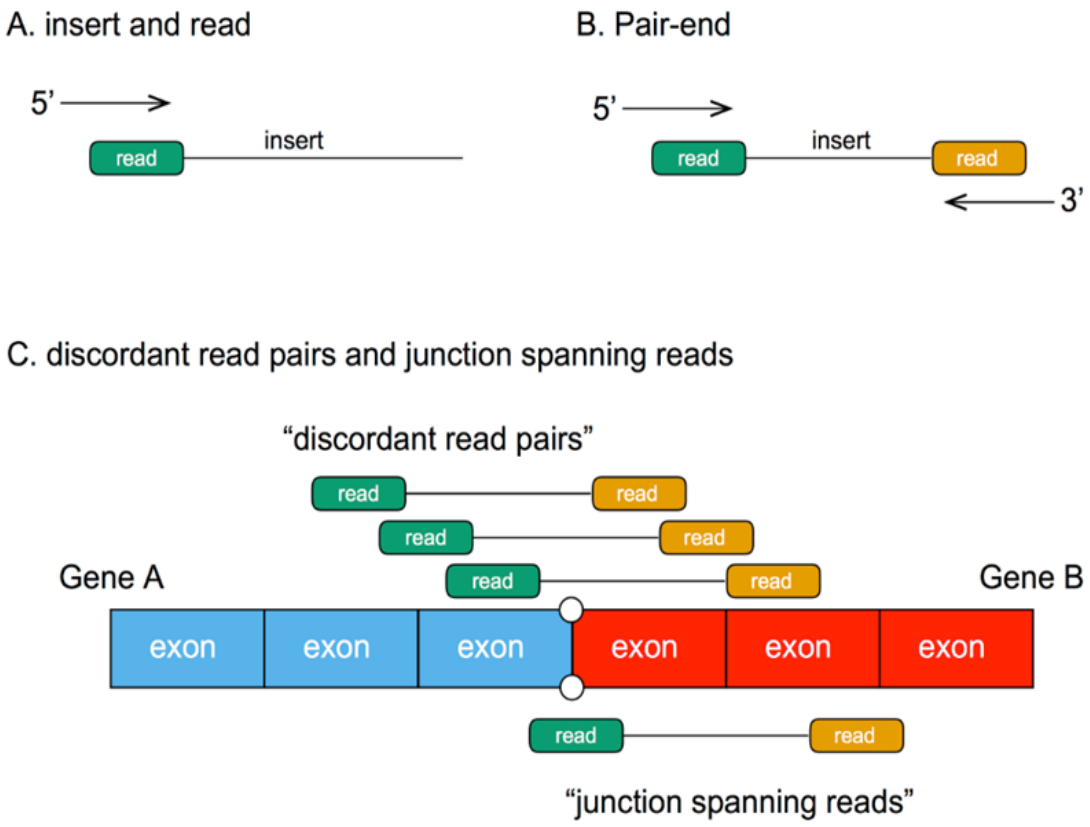


Figure 3. Detection of Fusion Transcripts using PRADA Algorithm. (A) Examples of sequencing: Single-end and (B) Paired-end. (C) The PRADA algorithm identifies fusion transcripts by detecting discordant read pairs and appraising apparent fusion junction spanning reads[25,26].

Differential Gene Expression

The gene expression profiles from the RNA-seq data were analyzed using DESeq2[27]. The differential gene expression analysis was performed on the raw data, as per recommendations of the[27] authors, and the cutoffs for the main parameters, fold change, and adjusted p-value, were modified based on the different analyses [see Appendix for R code]. Differentially expressed genes were grouped using complete linkage hierarchal clustering and visualized using pheatmap[28] to potentially reveal expression patterns associated with fusions.

Pathway and Interaction Network Analysis

To determine functional patterns of differentially expressed genes, we ran gene set enrichment analysis (GSEA)[29], a computational method that compares curated gene lists from MSigDB (Molecular Signatures Database) to experimental datasets[29]. The primary result of GSEA is the enrichment score (ES), showing the amount by which a predefined gene set is represented at the top or bottom of a list of the genes ranked by differential expression[29] [Fig. 4]. A positive score will mean that genes over-represented in that gene set are upregulated (top of the ranked list) in your dataset and a negative that genes over-represented in that gene set are downregulated (bottom of the ranked list)[29].

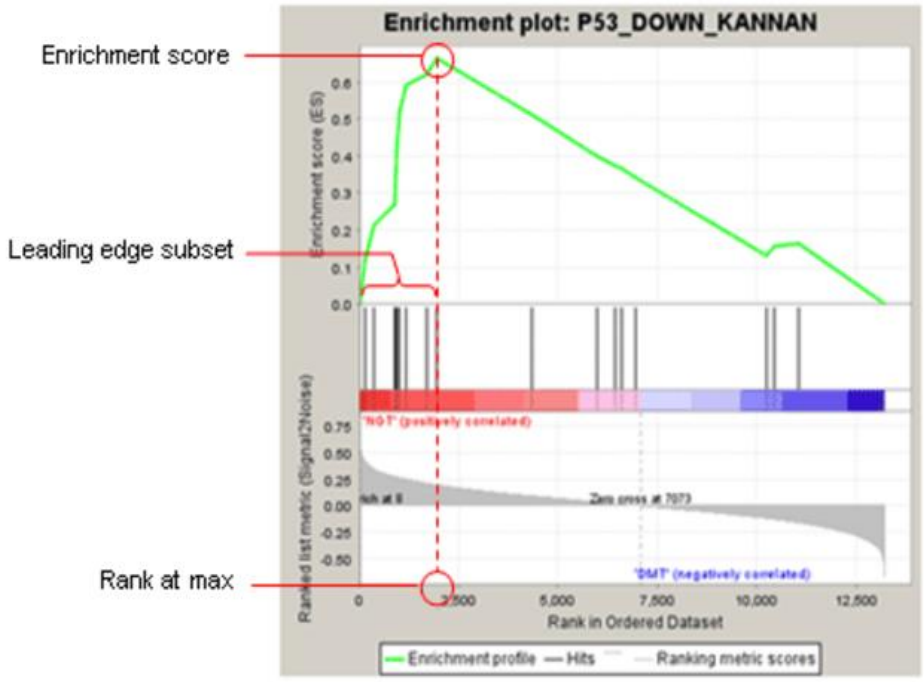


Figure 4. Example GSEA Result. GSEA example showing the enrichment score (ES), showing the amount by which a predefined gene set is represented at the top or bottom of a ranked list of genes. This is calculated by walking down the ranked list of genes, increasing and decreasing a running-sum statistic when a gene either is or is not in one of those curated lists[29].

To determine the interaction networks we used a software package called biomaRt[30], that accesses BioGRID[13], a curated database of publications for protein and genetic interactions. Using this tool, we found the known interaction partners of our differentially expressed genes and visualized them using Cytoscape[19]. Cytoscape is an interactive visualization tool that allows the user to incorporate interaction network information and gene expression profiles[19]. Finally, effected pathways were identified using the list of differentially expressed genes and their interaction partners to query three pathway databases: Reactome[17], KEGG Pathway[15,16], and Gene Ontology (GO)[18].

Exploratory Analysis - Clinical Prediction Models

We analyzed clinical metadata derived from the LUAD patient dataset to determine the effect of fusion status on patient survival. Our initial set contained 594 patients, and after filtering for missing information there were 516 patients. We fit a logistic regression model to determine if fusion status would better predict survival of lung cancer patients.

$$logit(p) = b_0 + b_1X_1 \ldots + b_nX_n$$

where p is the probability of a characteristic, the logit transformation is defined as the log odds:

$$odds = \frac{p}{1-p} = \frac{P(Presence\ of\ Characteristic)}{P(Absence\ of\ Characteristic)}$$

and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Based on the number meaningful features available we initially included five clinical variables in our model: fused/non-fused, age, tissue site, sample type, days to event, and tumor stage. We used k-fold cross-validation to partition the dataset, and then implemented step-wise regression on the training set to build the model, an algorithmic process that determines the feature set based on the Akaike information criterion (AIC). AIC estimates the quality of each model relative to all other potential models for a given dataset. We assessed the quality of our model against the iteratively partitioned test dataset. We also tested the model using an analysis of variance (ANOVA). Here the difference between the null and residual deviance reveals the degree to which a model explains the data in comparison to the null model. Finally, we compared whether the inclusion of fusion status in the model increased prediction. We tested the model with and without fusion status and calculated the prediction accuracy, ROC curves and AUC for both scenarios.

RESULTS

The Cancer Genome Atlas & Fusion Detection

The PRADA[25] algorithm identified many fusions, both novel and known, within the LUAD RNA-seq dataset from TCGA. PRADA[25] also reveals the presence of gene fusions in matched normal tissue samples, suggesting many fusions do not have an impact on tumorigenesis. In our analysis, we used all normal tissue samples, regardless of the presence of fusions. To determine how fusion tumors are distinct from healthy tissue, we compared EML4-ALK fusions against normal samples. To further identify the differences between fusion and non-fusion tumors, we compared EML4-ALK positive tumors against those without the fusion. [Fig. 5].

Status		Tumor	Normal	Total
Fusions		370	19	389
	EML4-ALK Fusions*	5	0	5
Non-Fused		146	40	186
Total		521	59	580

Figure 5. Characterization of Samples within the LUAD Dataset. This table shows how many unique patient samples there are within each sample type. The total number of unique patient samples is 580, the original dataset however was 594 samples due to the inclusion of some technical replicates. *EML4-ALK fusions are a specific subset of fusions within the LUAD data.

Differential Gene Expression

EML4-ALK Fusion Tumors Compared to Normal Samples

Our first analysis compared the RNA-seq data of 59 normal patient samples to 5 EML4-ALK positive samples. Differential gene expression analysis using DESeq2[27] found 894 differentially expressed genes with more than a 4-fold change and an adjusted p-value of less than 0.001. Unsupervised hierarchal clustering of the 894 genes, using normalized $\log_2$ transformed counts showed distinct patterns between fusions and non-fusions [Fig. 6].

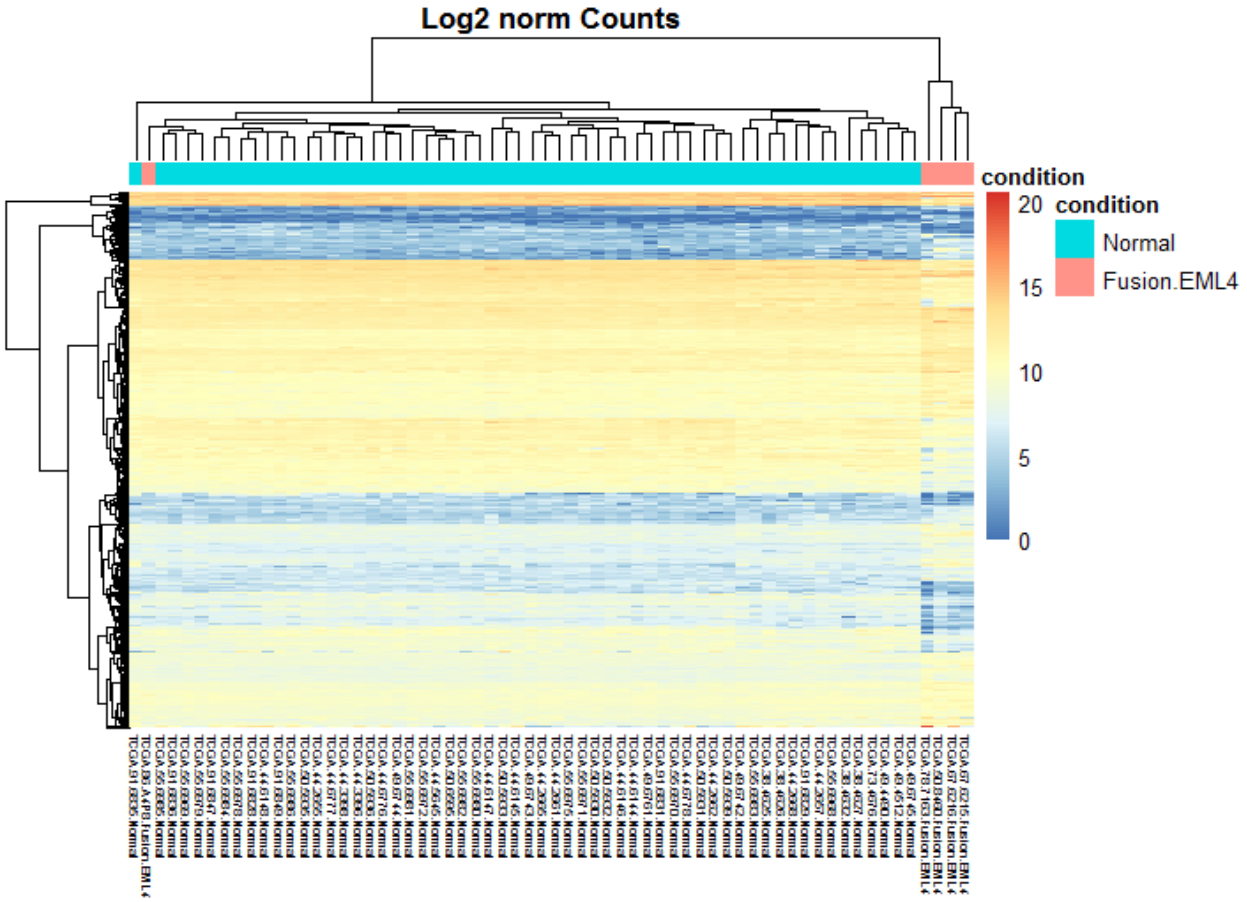


Figure 6. Heatmap of Fusion Containing Tumors Compared to Normal Samples. Unsupervised clustering was done on 894 differentially expressed genes between 59 normal tissue samples and 5 EML4-ALK fusion samples. The unsupervised clustering shows that four fusion samples on the right have a distinct profile when compared to the normal samples.

In fact, hierarchal clustering of the patient samples grouped all but one of the fusions as distinctly separate from the normal samples. A closer inspection of the top differentially expressed genes reveals several cohesive biological functions [Fig. 7]. Heat shock proteins (e.g. Hsp90AB1) are shown to be up-regulated and hemoglobin beta (HBB) is down-regulated.

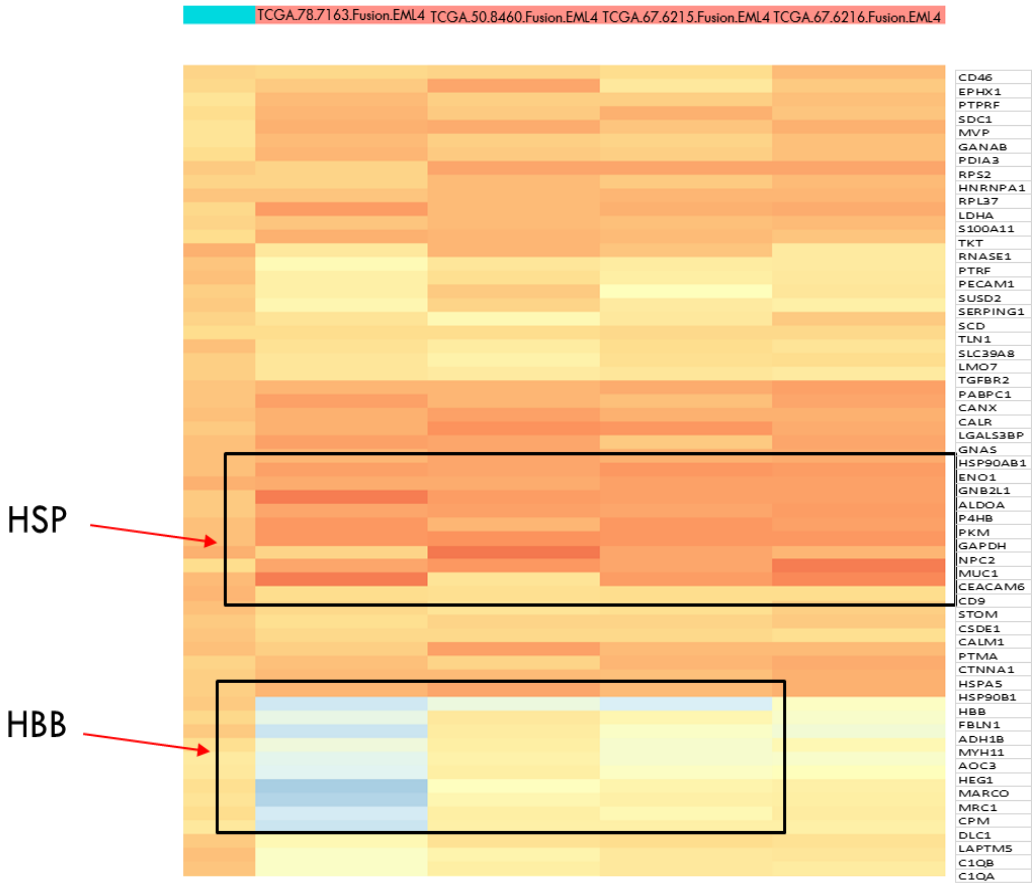


Figure 7. Differential Expression of Heat Shock Proteins and Hemoglobin in EML4-ALK Fusion Samples. Unsupervised clustering was done on 894 differentially expressed genes between 59 normal tissue samples and 5 EML4-ALK fusion samples. These are two highlighted biological functions that were found to be differentially expressed in four of fusion samples when compared to the normal samples.

EML4-ALK Fusion Tumors Compared to Non-Fused Tumors

The second analysis compared EML4-ALK fusion positive tumors to fusion-negative tumors. There were 146 patients with no fusions in their tumors and 5 patients that had the EML4-ALK fusion in their tumor sample. We classified differentially expressed genes as those that had more than a 2-fold change and an adjusted p-value of less than 0.05. The 114 identified genes were clustered in a similar manner using normalized and $\log_2$ transformed counts [Fig.8]. Perhaps more striking that our previous comparison, clustering could group all fusion tumors as a top-level group distinct from all other normal samples except for one normal sample. These results suggest a substantive biological distinction between even fusions positive tumors and fusion negative tumors.

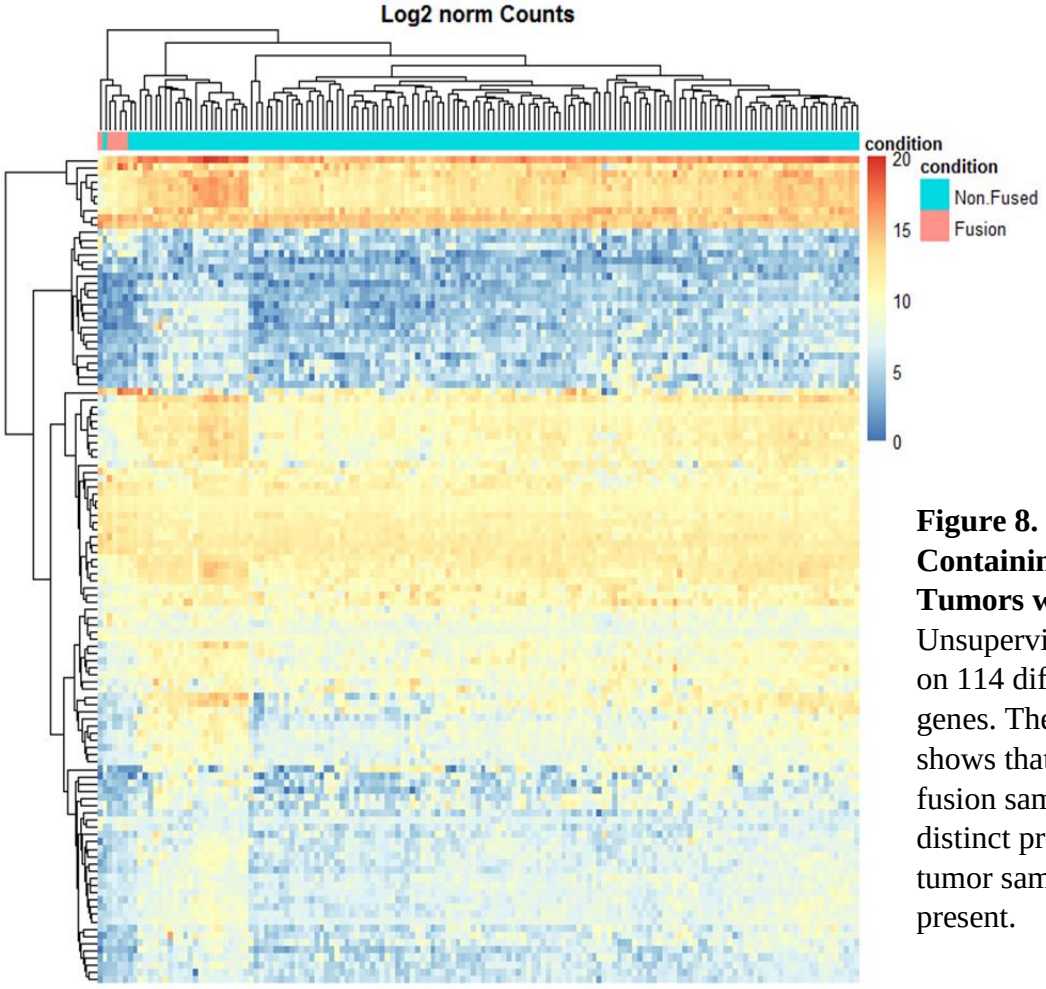


Figure 8. Heatmap of Fusion Containing Tumors Compared to Tumors without any Fusions. Unsupervised clustering was done on 114 differentially expressed genes. The unsupervised clustering shows that the five EML4-ALK fusion samples on the left have a distinct profile when compared to tumor samples without any fusions present.

Several genes highlight the differences between tumors with fusions and those without [Fig. 9]. CHI3L1, and to no surprise, ALK stand out in this expression profile, and show a clear increase in expression with the fused patient samples compared to the non-fused. ALK is the oncogene in the EML4-ALK fusion and is the most differentially expressed gene out of the entire 114 gene list and had an incredible 14.7-fold-change and an adjusted p-value of 1.23×10^{-7}.

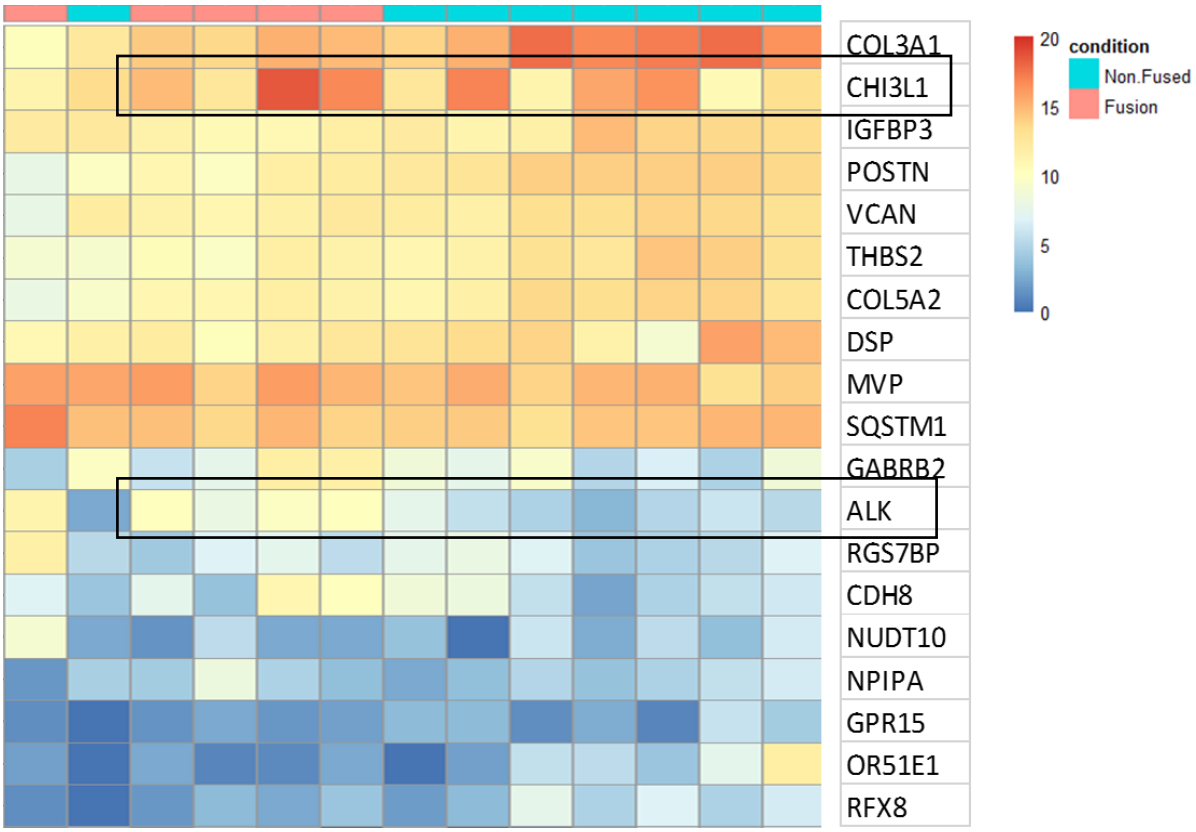


Figure 9. Differential Expression of CHI3L1 and ALK in EML4-ALK Fusion Samples. Unsupervised clustering was done on 114 differentially expressed genes between 146 non-fused tumor samples and 5 EML4-ALK fusion samples. CHI3L1 and ALK that were found to be among the most differentially expressed genes in five fusion samples when compared to samples without any fusions.

Pathway and Interaction Network Analysis

EML4-ALK Fusion Tumors Compared to Normal Samples

To analyze the over-arching functional categories of the differential expressed genes, we GSEA used on the 894 significant genes and their normalized count data to determine which predefined gene set had the highest enrichment score (ES) and the most correlated genes. We found that the ribonucleoside diphosphate metabolic process and the highest ES for the fusion phenotype, indicating a strong representation of that gene set in the upregulated genes of the fusion phenotype [Fig. 10]. Other enriched pathways include the nucleoside diphosphate metabolic process, nucleotide phosphorylation, and ADP metabolic process, all of which are involved in DNA and RNA synthesis, regulation, and repair.[31,32]

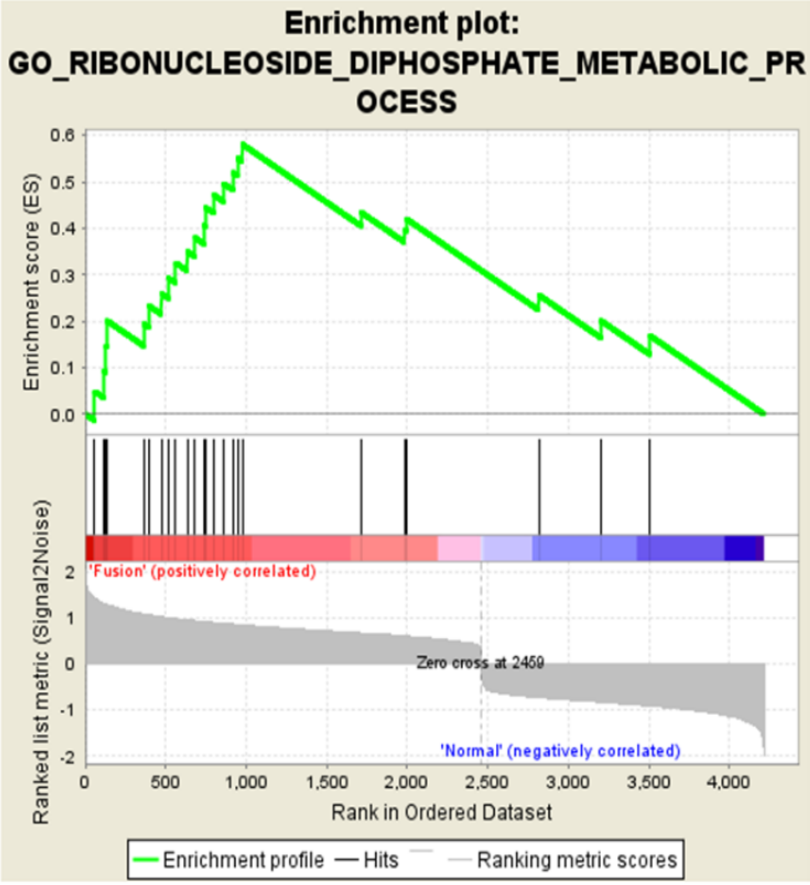


Figure 10. Gene Set Enrichment Analysis (GSEA) on 894 Differentially Expressed Genes Between Normal and EML4-ALK Fusion Samples. GSEA results showing the enrichment score (ES), showing the amount by which this ribonucleoside diphosphate metabolic process gene set is represented in the 894-ranked list of genes. This positive ES means that there is an over-representation of upregulated genes in the fusion phenotype.

To explore enriched gene sets and pathways beyond GSEA, we utilized the KEGG Pathway[15,16] and GO[18] databases to see if the 114 differentially expressed genes were associated with any specific pathways. We found that 81 genes out of the list that were associated with mitogen-activated protein kinase (MAPK) family signaling cascades, and 87 genes that were associated with receptor tyrosine-protein kinase erbB4 (ERBB4) signaling pathways. Both pathways involve tyrosine kinases which are integral to regulation of cell growth, differentiation and survival.

EML4-ALK Fusion Tumors Compared to Non-Fused Tumors

We conducted a similar gene set enrichment on the differentially expressed genes between the ELL4-ALK fusion positive tumors and fusion free tumors. The analysis showed positive enrichment for band 22 on chromosome 16 [Fig. 11]. The specific genes at this locus - TERF2, CDH8, TK2, DDX19A, LRRC29 and HP – were all found to be positively correlated with the EML4-ALK fusion type.

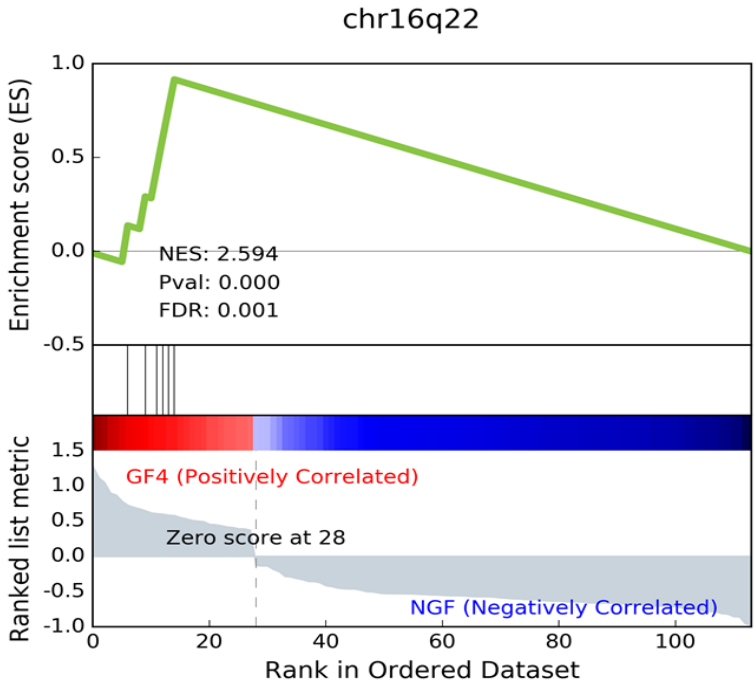


Figure 11. Gene Set Enrichment Analysis (GSEA) on 114 Differentially Expressed Genes Between EML4-ALK Fusion Samples and Non-Fused Tumor Samples. GSEA results showing the enrichment score (ES), showing the amount by which chromosome 16, band 22 gene set is represented in the 114-ranked list of genes. This positive ES means that there is an over-representation of upregulated genes in the fusion phenotype.

Due to the small number of differentially expressed genes in this analysis, querying KEGG[15,16] and GO[18] with our gene list did not reveal any major pathways, and returned a widely-varied list processes. However, the smaller gene list allowed us to explore the gene interaction network by identifying interaction partners of all the genes. The previous gene list of 894 genes proved too large to cohesively analyze through this method.

The 114 differentially expressed genes were evaluated by biomaRt[30] to find interaction partners associated with them from the repository bioGRID[13]. There were 291 unique genes found to interaction with the list of genes comparing fusion tumor samples to non-fused tumor samples. Some genes were also duplicated due to interaction with itself and other genes already within the differentially expressed gene list.

To visualize differentially expressed genes and their interaction network, the 114 genes and their interaction partners were plugged into Cytoscape[19] [Fig. 12]. While there are some isolated interactions, most genes form a single network. Within the main network, there were two main clusters of genes, suggesting broader and more important role for these hubs [Fig.13].

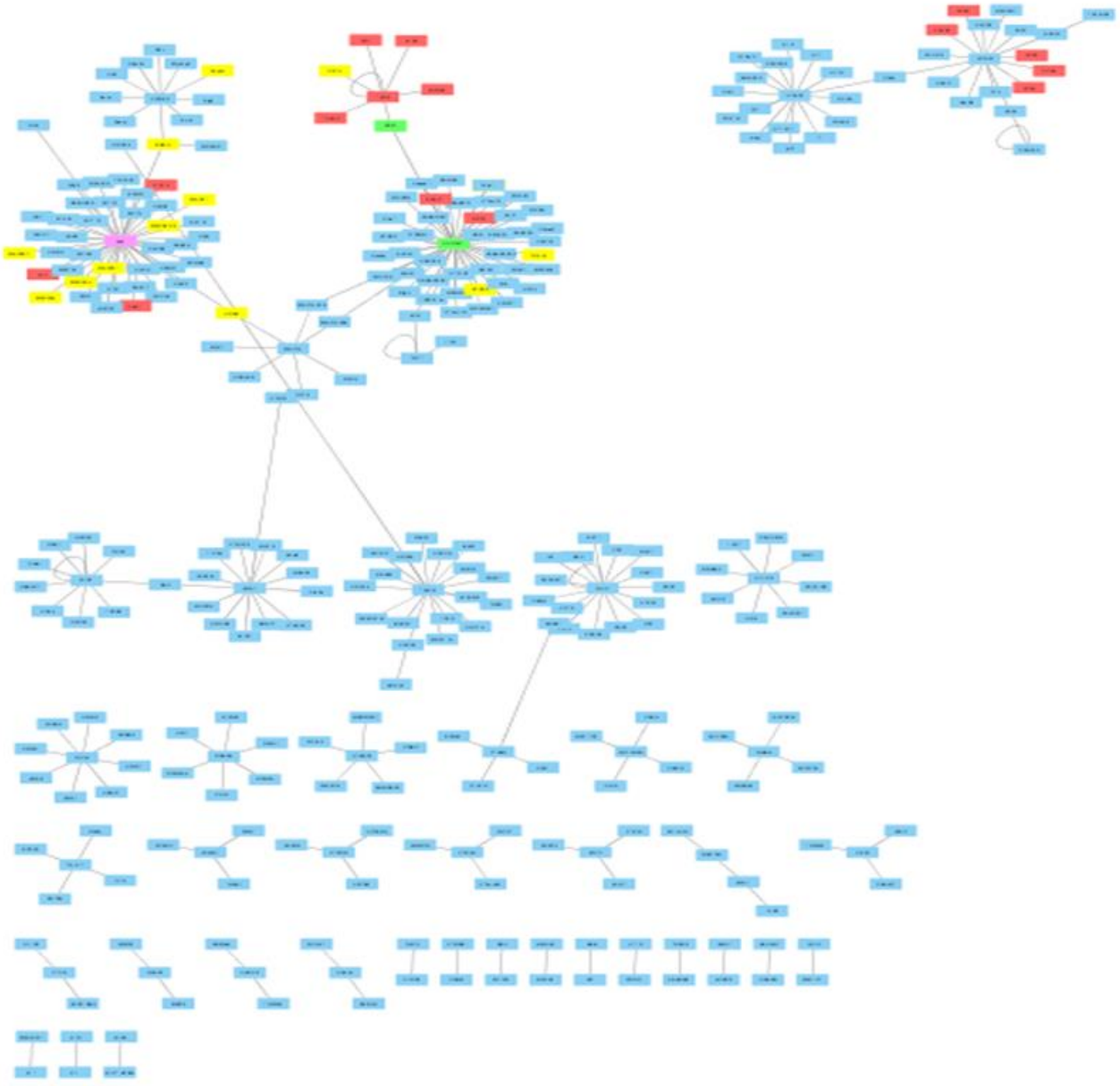
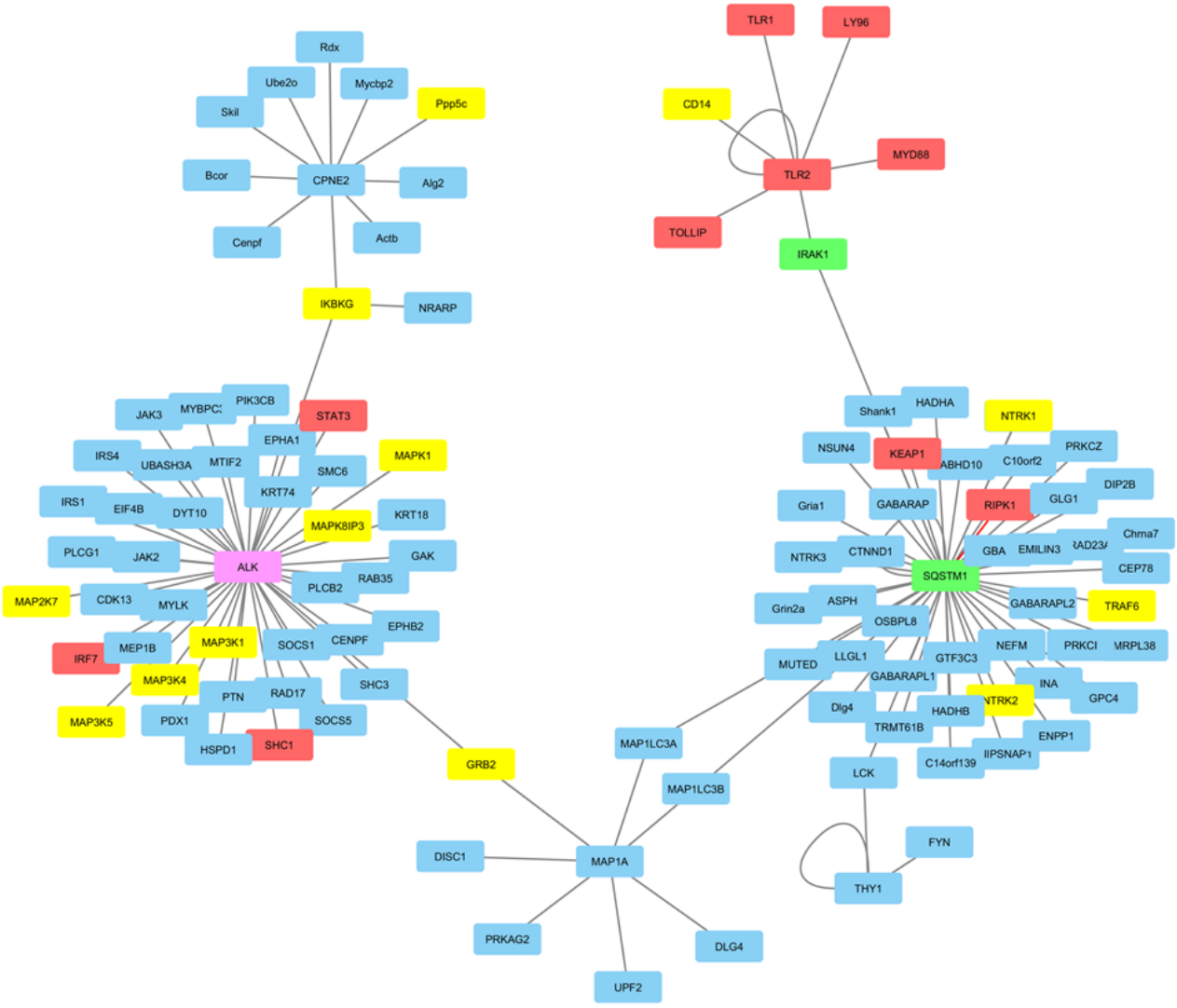


Figure 12. Interaction Network for Comparing EML4-ALK Fusion Samples to Non-Fused Tumor Samples: All Interactions. Using the 114 differentially expressed gene list and known interaction partners we visualized the interaction network in Cytoscape[19]. While there are some isolated interactions, most genes forma a single network.



Annotations:

- **Red** = Immune Response
- **Yellow** = MAPK
- **Green** = NF-κB
- **Pink** = ALK

Figure 13. Interaction Network for Comparing EML4-ALK Fusion Samples to Non-Fused Tumor Samples: Main Cluster. Using the 114 differentially expressed gene list and known interaction partners we visualized the interaction network in Cytoscape[19]. Most interactions are isolated in the overall network but this single network stood out with the most interactions. The color coding shows what pathways and processes the genes are associated with. The genes that are colored are connected to known functional networks, and ALK is featured prominently at the center of a cluster of genes that are involved in the immune response pathways and the MAPK pathways.

In this visualization, ALK and SQSTM1 are featured prominently at the center of these clusters, and are identified as being involved in immune response pathways, MAPK pathways, and NF-κB pathways. We also observed a smaller network that included many inflammatory and immune response genes, like chemokine ligands (CCL), and VCAN. However, aside from the main network cluster, almost all other genes had single interactions.

Exploratory Analysis - Clinical Prediction Models

In this exploratory analysis, we examined if fusion genes have an additional impact on patient survival. We extracted the paired clinical data from the TCGA LUAD dataset, and found the mean survival time within this cohort to be 2.2 years, a mean follow-up time of 2.4 years, and there a max follow-up time of 19.9 years.

The best logical model identified using k-fold cross validation consisted of the following features: fusion status, patient age, days to event, and tumor stage. Validating the model on our test dataset, we found that the cancer stage is not a statistically significant feature in the prediction of patient survival. Our findings suggest that the model would have performed almost as well without tumor stage included as a feature.

The deviance table generated through ANOVA was used to assess the fit of the model compared to the null model, i.e. the hypothesis that none of the features are useful predictors of survival [Fig. 14]. A lower value for the residual deviance indicates a stronger prediction model. All the model variables were shown to significantly reduce the residual deviance. A less significant p-value here indicates a variable is not as essential for explaining variance in the data.

Feature	Degrees of freedom	Deviance	Residual Degrees of Freedom	Residual Deviance	Pr(>Chi)
Null			**392**	**458.22**	
Age	**1**	**3.97**	**391**	**454.25**	**0.04632**
Days to Event	**1**	**47.418**	**390**	**406.83**	**5.73E-12**
Tumor Stage	**7**	**46.227**	**383**	**360.61**	**7.90E-08**
Fusion Status	**1**	**4.161**	**382**	**356.45**	**0.04136**

Figure 14. ANOVA: Analysis of Deviance Table. This table assesses the model strength by comparing the null deviance and the residual deviance for the inclusion of each feature to the null model. A lower the value for the residual deviance indicates a stronger prediction model.

Including fusion status improved the prediction of survival for LUAD patients from an accuracy of 69% to 75%. In addition, the ROC curve shows the model has high sensitivity and specificity [Fig. 15] with an area under the curve (AUC) of 0.857. Based on this preliminary model, we found that including fusion status slightly increases the accuracy and the predictability of survival in patients with lung adenocarcinoma.

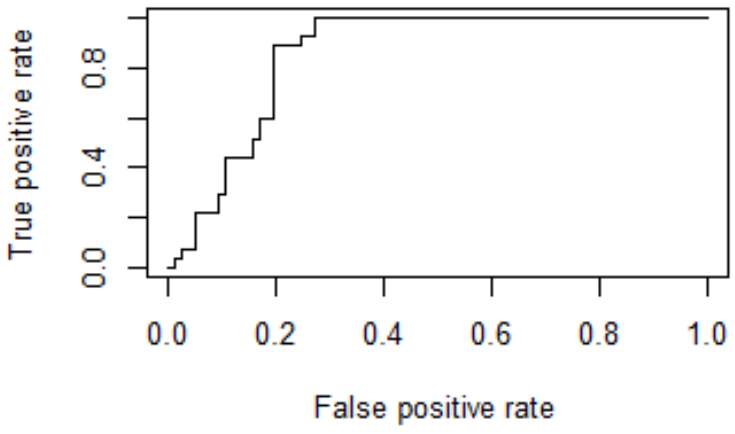


Figure 15. Receiver Operating Characteristic (ROC) curve. The prediction model was fit on the test data set and an ROC curve was used to evaluate its sensitivity and specificity. The true positive rate (sensitivity) is plotted against the false positive rate (100-specificity).

DISCUSSION

Many studies have shown that EML4-ALK drives tumorigenesis, making it a prime candidate for a proof-of-principle study in this thesis. We strived to explore the impact of this fusion beyond its mere presence in LUAD, and determine the impact it has on global gene expression patterns, pathways, and gene interaction networks, to demonstrate an informatics approach to understanding fusion driven biological changes. We analyzed the global expression profiles of the 594 LUAD tumor samples and 59 normal tissue samples found in TCGA database. To understand the differences between healthy tissue and fusion positive tumor biology, we compared the RNA-seq data of EML4-ALK tumors against normal tissue, and to more critically understand the difference between fusion-positive tumors and non-fusion tumors, we compared ELM4-ALK positive tumors against those with no detectable fusions.

How do Gene Fusions Change the Expression of Pathways?

The results of the unsupervised clustering and GSEA for both comparisons, show clear differences in the expression profiles and the pathways of differentially expressed genes. In comparing normal to EML4-ALK samples there was a striking difference between the 5 fused tumor samples and the 59 normal samples [Fig. 6], and clustering revealed up-regulation of heat shock proteins and down-regulation of hemoglobin beta [Fig. 7].

The upregulation of heat shock proteins has been shown to be indicative of poor prognosis in some cancers, in particular - gastric, prostate, and breast cancer[33]. Heat shock proteins are also known to elicit a response by the immune system and have been implicated in tumor proliferation, differentiation, and invasion[1,33,34]. Specifically, Hsp90 and its partners are chaperone proteins that function to maintain other vital tumor-promoting client proteins [33,34].

Decreased HBB has been identified in anaplastic thyroid cancer (ATC)[35,36], and malignancies in general have been associated with global dips in hemoglobin level[35]. HBB is located on chromosome 11 band 15, a known tumor suppressing locus that is susceptible to genomic instability and loss of heterozygosity (LOH), where a wild-type allele is lost and the cell is left with a disease-causing allele[36].

The GSEA results comparing fusions against normal tissue samples showed that a high number of differentially expressed genes are involved in the ribonucleoside diphosphate metabolic process [Fig. 10]. Ribonucleoside diphosphate reductase is a key enzyme in formation of deoxyribonucleotides and its upregulation would seem to indicate higher rates of replication, a known feature of cancerous cells[31,32]. Most of the identified pathways appear to be involved in DNA/RNA synthesis, regulation, and repair.

Comparing the fusion-free tumor samples to EML4-ALK fusions also showed a distinct profile for the 5 fused tumor samples [Fig. 8]. The gene CHI3L1 showed significant expression changes, and has been characterized as a glycoprotein associated with inflammation processes in response to infections[37]. It has been associated with poor prognosis in breast cancer, and found to be up-regulated in physically large tumors, although its function is not completely clear[38]. ALK, the namesake member of the oncogenic EML4-ALK fusion is also clearly up-regulated in fusion-positive patient samples compared to non-fused samples. Out of all 114 differentially expressed genes, it is both the most significant and most highly up-regulated gene. The significantly large increase in expression of ALK is a result of translocation to the EML4 promoter, and drives the oncogenic properties of the EML4-ALK fusion gene[22]. It is thought that EML4 is incidentally involved in the oncogenic properties of the fusion appropriation of its promotor, and is otherwise characterized as an echinoderm microtubule-associated protein involved in microtubule

formation[22]. ALK is an anaplastic lymphoma kinase, which is a receptor tyrosine kinase, these kinases are cell surface receptors that bind and respond to growth factors play a huge role in regulation of cell growth, differentiation and cell survival[39]. The EML4-ALK fusion causes dimerization of the tyrosine kinase without a ligand binding to it leading to the unregulated over-expression of the protein[22]. This in turn activates many pathways, including MAPK, JAK/STAT pathway, the PI3K/Akt pathway, which are involved in cell proliferation, differentiation, and cell survival, leading to its oncogenic power[22].

GSEA on the 114-gene list enriched for chromosome 16 band 22 [Fig. 10]. This otherwise unremarkable locus has been found to be associated with other cancer types[40]. Chromosome 16 is unusual compared to other chromosomes in that it is enriched in repetitive sequences[41]. It is believed that these repeat sequences undergo more frequent mutation events, occasionally resulting in chromosomal rearrangements. However, this chromosome does not coincide with either partner of the EML4-ALK fusion[41]. Nevertheless, the specific genes at this locus - TERF2, CDH8, TK2, DDX19A, LRRC29 and HP – are all upregulated in ELM4-ALK fusions. TERF2 codes for a telomere repair protein, which may hint at its role in mitigating genomic degradation during the rapid division of cancer cells[42].

Our results indicate the expression of different pathways changes in samples with the EML4-ALK fusion genotype in comparison to both normal and fusion-free tumor tissue. These differences lead to ELM4-ALK fusions clustering distinctly from other samples. Looking deeper into the differentially expressed genes, there are different pathways and processes that are enriched based on these different phenotypes. The fusion phenotype is enriched in pathways that regulate the synthesis of DNA and its repair, and in locations that are highly repetitive and unstable. It also

relates to increasing genomic instability, as instability can both cause and result in shortening telomeres, which can be in part counteracted by the function of TERF2[42].

How do Gene Fusions Change the Interaction of Pathways?

While the introduction of gene fusions causes dysregulation of pathway components, it is unclear whether impacted pathways alter their functions and interact with new components. Comparing the EML4-ALK fusion tumors to normal samples, a significant number of differentially expressed genes were associated with MAPK family signaling cascades and ERBB4 signaling pathways. The MAPK pathway includes many signaling molecules like ERK, Ras, and Raf, which are normally activated by growth factors binding to the receptor tyrosine kinases[43]. Under normal conditions, MAPK plays a vital role in regulation of cellular growth and proliferation[43]. However, an increase in expression of MAPK genes would imply the dysregulation of the pathway, and activation of signaling has caused an increased or uncontrolled cell proliferation[43]. ERBB4, also known as HER4, is a member of the EGFR subfamily of receptor tyrosine kinases and once activated induces a variety of cellular functions including cell differentiation[44]. Studies have shown that mutations to this pathway cause uncontrolled signaling, have been found in other cancer types, including non-small cell lung cancer[44].

Cell line studies in non-small cell lung cancer with EML4-ALK fusions have shown that the fusion drives the phosphorylation of Akt and ERK signaling molecules[45]. Akt is involved in the survival-associated PI3K-Akt pathway and ERK is in the MAPK pathways, which both have longstanding associations with many cancer types because of their involvement in many cellular functions that control growth, proliferation and differentiation[22,43]. While there is not a definitive answer, EML4-ALK might be upstream of STAT3 phosphorylation[22]. This would mean the

activation of the JAK-STAT pathway, that is involved in transcription and expression of genes involved in apoptosis, immune response and proliferation[22].

Interaction network analysis and Cytoscape[19] visualizations of the comparison of EML4-ALK fusion and fusion-free tumors showed a formation of a large interconnected network with a limited number of isolated clusters [Fig. 12 & 13]. It also showed ALK was centrally located in the main hub of the large network, surrounded by genes involved in immune response, MAPK and NF-κB. Although, it is doubtful that the immune response is effected in some way due to the gene fusion, novel antigens due to genetic rearrangements may cause inflammation or recruit the immune system[4,5]. In fact, expression of NF-κB, a transcription factor, is typically found in inflammatory and immune responses [46]. In addition, Meylan et. al demonstrated the role of the transcription factor in tumor development by showing that loss of oncogenes p53 and upregulation of $KRAS^{G12D}$ resulted in NF-κB pathway activation[46].

Within this study, we found many of the same signaling pathways: MAPK, JAK-STAT pathway, the PI3K-Akt, arising in different parts of our analyses. The interactions of these critical signaling pathways suggest an underlying coregulation, and ultimately dysregulation due to the EML4-ALK fusion.

FUTURE WORK

While we gained crucial insight into fusion biology, there were several key limitations in our analyses. First, our focus was on EML4-ALK, though other known and novel fusions have been described in the literature. During our literature review of well characterized fusions, we curated a list of 167 gene fusions, their functions, and pathway information. Of them, we believe the BCR-ABL1 and CD74-ROS1 fusions are promising for a follow-up study because they are

well annotated genes individually and fused. Each of the genes in the fused pair are involved in different pathways and have very different functions.

Second, the TCGA dataset only had a small number (N=5) of patients with EML4-ALK, though this fusion was the most abundant. It suggests that fusions both vary in partners, and are non-homogenous across patients and cancers. To overcome this, we could incorporate multiple fusions as well as other cancers, drawing from the larger TCGA database. In addition, we could leverage the Genotype-Tissue Expression (GTEx) project to gather additional tissue specific gene expression data.

Third, we did not explore overlaying expression data on known pathways, which would shed light on how precisely pathways are changed by fusions. Such an analysis might reveal how changes in metabolic flux translate into biological significance, and provide more insight into how excess products from one pathway go on to affect other pathways.

This research contributes to the fields of cancer biology and bioinformatics by taking preliminary steps in analyzing gene fusions and their impacts on pathways. We found many common themes in differentially expressed genes, and their effect on critical cellular functions. Our approach could be generalizable and scalable for analyzing many other gene fusions quickly and efficiently. Translating this analytical workflow into an automated cloud driven pipeline would make it accessible to the broader research and clinical community. And while our approach only utilizes RNA-seq and clinical data, other data sources such as proteomics and metabolomic data could be incorporated to further strengthen the system biology approach we take.

CONCLUSIONS

Advancements in next-generation sequencing technology have made gene fusions easier to detect. Beyond identification, only a handful of studies have been performed to understand the function of specific fusions. A comprehensive approach, leveraging existing tools and biological resource databases, is necessary to further understand the complexities of fusion driven changes. In this research, we study the impact of gene fusions on biological pathways in the context of cancer, by tying together current tools to gain new insights. We used the EML4-ALK as the model fusion in studying the lung adenocarcinoma dataset in TCGA. We demonstrate differences of gene expression and pathway regulation between healthy tissue, fusion-free tumors, and fusion-positive tumors. We find that ELM4-ALK fusions have discrete expression profiles, specific pathways that are activated or dysregulated due to the fusion and there are clear interaction networks that form around common biological responses. Clearly fusions have an impact not only on the pathways they are involved in but also the pathways that genes they interact with are involved in. Our approach demonstrates the potential of using large multi-omics datasets to fundamentally reanalyze the biological pathway changes propagated by gene fusions.

BIBLIOGRAPHY

1. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013.
2. Editor G, Shen Z. Genomic Instability and Cancer Metastasis. 2015;20:1-3. doi:10.1007/978-3-319-12136-9.
3. Davis RJ, Barr FG. Fusion genes resulting from alternative chromosomal translocations are overexpressed by gene-specific mechanisms in alveolar rhabdomyosarcoma. *Proc Natl Acad Sci U S A*. 1997;94(15):8047-8051. doi:10.1073/pnas.94.15.8047.
4. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer*. 2015;15(6):371-381. doi:10.1038/nrc3947.
5. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*. 2007;7(April):233-245. doi:10.1038/nrc2091.
6. Zhang J, White NM, Schmidt HK, et al. INTEGRATE: Gene fusion discovery using whole genome and transcriptome data. *Genome Res*. 2015:1-11. doi:10.1101/gr.186114.114.
7. McPherson A, Hormozdiari F, Zayed A, et al. Defuse: An algorithm for gene fusion discovery in tumor rna-seq data. *PLoS Comput Biol*. 2011;7(5). doi:10.1371/journal.pcbi.1001138.
8. Latysheva NS, Babu MM. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res*. 2016;44(10):4487-4503. doi:10.1093/nar/gkw282.
9. Latysheva NS, Oates ME, Maddox L, et al. Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer. *Mol Cell*. 2016;63(4):579-592. doi:10.1016/j.molcel.2016.07.008.
10. NIH. Biological Pathways. https://www.genome.gov/27530687/. Published 2015.
11. Platia J, Bucura O, Khosravi-Far R. Apoptotic cell signaling in cancer progression and therapy. *Integr Biol*. 2011;3(4):279-296. doi:10.1039/c0ib00144a.Apoptotic.
12. Gene I, Networks I. Statistical Human Genetics. 2012;850:483-494. doi:10.1007/978-1-61779-555-8.
13. Chatr-Aryamontri A, Oughtred R, Boucher L, et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res*. 2017;45(D1):D369-D379. doi:10.1093/nar/gkw1102.
14. Wu C-C, Kannan K, Lin S, Yen L, Milosavljevic A. Identification of cancer fusion drivers using network fusion centrality. *Bioinformatics*. 2013;29(9):1174-1181. doi:10.1093/bioinformatics/btt131.
15. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(D1):D353-D361.

doi:10.1093/nar/gkw1092.

16. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 1999;27(1):29-34. doi:10.1093/nar/27.1.29.

17. Fabregat A, Sidiropoulos K, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2016;44(D1):D481-D487. doi:10.1093/nar/gkv1351.

18. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: Tool for The Unification of Biology. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556.

19. Christmas, Rowan; Avila-Campillo, Iliana; Bolouri, Hamid; Schwikowski, Benno; Anderson, Mark; Kelley, Ryan; Landys, Nerius; Workman, Chris; Ideker, Trey; Cerami, Ethan; Sheridan, Rob; Bader, Gary D.; Sander C. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Am Assoc Cancer Res Educ B*. 2005;(Karp 2001):12-16. doi:10.1101/gr.1239303.metabolite.

20. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol*. 2011;12(8):15. doi:10.1186/gb-2011-12-8-r72.

21. Hanahan D. The Hallmarks of Cancer. *Cell*. 2000;100(1):57-70. doi:10.1016/S0092-8674(00)81683-9.

22. Bayliss R, Choi J, Fennell DA, Fry AM, Richards MW. Molecular mechanisms that underpin EML4-ALK driven cancers and their response to targeted drugs. *Cell Mol Life Sci*. 2016;73(6):1209-1224. doi:10.1007/s00018-015-2117-6.

23. The Cancer Genome Atlas (TCGA). https://cancergenome.nih.gov/.

24. Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci Rep*. 2016;6(1):21597. doi:10.1038/srep21597.

25. Torres-garcía W, Zheng S, Sivachenko A, et al. Application Notes PRADA : Pipeline for RNA sequencing Data Analysis. *Bioinforma Adv access*. 2014:4-5. doi:10.1038/nature12222.Verhaak.

26. Yoshihara K, Wang Q, Torres-Garcia W, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*. 2015;34(37):4845-4854. doi:10.1038/onc.2014.406.

27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.

28. Kolde R. pheatmap: Pretty Heatmaps. 2015. https://cran.r-project.org/package=pheatmap.

29. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102.

30. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*.

2005;21(16):3439-3440. doi:10.1093/bioinformatics/bti525.

31. Elledge SJ, Zhou Z, Allen JB. Ribonucleotide reductase: regulation, regulation, regulation. *Trends Biochem Sci*. 1992;17(3):119-123. doi:10.1016/0968-0004(92)90249-9.

32. Herrick J, Sclavi B. Ribonucleotide reductase and the regulation of DNA replication: An old story and an ancient heritage. *Mol Microbiol*. 2007;63(1):22-34. doi:10.1111/j.1365-2958.2006.05493.x.

33. Ciocca DR, Calderwood SK. Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. *Cell Stress Chaperones*. 2005;10(2):86. doi:10.1379/CSC-99r.1.

34. Koga F, Kihara K, Neckers L. Inhibition of cancer invasion and metastasis by targeting the molecular chaperone heat-shock protein 90. *Anticancer Res*. 2009;29(3):797-807. http://www.ncbi.nlm.nih.gov/pubmed/19414312.

35. Taylor A, Ph D, Pollack MA, Ph D. Hemoglobin Level and Tumor Growth*. 1942.

36. Onda M, Akaishi J, Asaka S, et al. Decreased expression of haemoglobin beta (HBB) gene in anaplastic thyroid cancer and recovery of its expression inhibits cell growth. *Br J Cancer*. 2005;92(12):2216-2224. doi:10.1038/sj.bjc.6602634.

37. Mehta P, Schwab DJ, Sengupta AM. NIH Public Access. *Changes*. 2012;29(6):997-1003. doi:10.1016/j.biotechadv.2011.08.021.Secreted.

38. Wan G, Xiang L, Sun X, Wang X, Li H. Elevated YKL-40 expression is associated with a poor prognosis in breast cancer patients. 2017;8(3):5382-5391.

39. Robinson DR, Wu Y-M, Lin S-F. The protein tyrosine kinase family of the human genome. *Oncogene*. 2000;19(49):5548-5557. doi:10.1038/sj.onc.1203957.

40. Mantripragada K, Caley M, Stephens P, et al. Telomerase Activity is a Biomarker for High Grade Malignant Peripheral Nerve Sheath Tumors in Neurofibromatosis Type 1 Individuals. *Genes Chromosomes Cancer*. 2008;47(April):238-246. doi:10.1002/gcc.

41. Mapping T. What's Different About Chromosome 16? *Los Alamos Sci*. 1992;(20):211-215.

42. Karlseder J, Hoke K, Mirzoeva OK, et al. The telomeric protein TRF2 binds the ATM Kinase and Can Inhibit the ATM-dependent DNA damage response. *PLoS Biol*. 2004;2(8). doi:10.1371/journal.pbio.0020240.

43. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. *Oncogene*. 2007;26(22):3279-3290. doi:10.1038/sj.onc.1210421.

44. Kurppa KJ, Denessiouk K, Johnson MS, Elenius K. Activating ERBB4 mutations in non-small cell lung cancer. *Oncogene*. 2016;35(10):1283-1291. doi:10.1038/onc.2015.185.

45. Petrochilos D, Shojaie A, Gennari J, Abernethy N. Using random walks to identify cancer-associated modules in expression data. *BioData Min*. 2013;6(1):17. doi:10.1186/1756-0381-6-17.

46. Meylan E, Dooley AL, Feldser DM, Shen L, Turk E, Ouyang C. Requirement for NF-κB signaling in a mouse model of lung adenocarcinoma. 2010;462(7269):104-107. doi:10.1038/nature08462.Requirement.

47. Leonard G. Gene Fusion Types.

APPENDIX

TCGA File Scrub Code

```
##Code to Read the Clinical data from the JSON file
library(jsonlite)
library(dplyr)
library(Hmisc)
library(data.table)
install.packages('Hmisc')
# reading in the json for biospecimen
tmp_biospecimen <- fromJSON('C:/Users/thood/Downloads/biospecimen.cart.2016-09-07T21-46-35.302439.json')
# rbindlist of biospecimen samples for inspecting sample_type
tmp_biospecimen_2 <- tmp_biospecimen$samples %>% rbindlist() %>% data.table()
tmp_biospecimen_2 %>% select(-portions) %>% describe()
tmp_biospecimen_2 %>% select(c(sample_id)) %>% unique() %>% arrange(sample_id) %>% table()
tmp_clin <- fromJSON('C:/Users/thood/Downloads/clinical.cart.2016-09-07T21-34-57.820511.json')
tmp_clin$case_id %>% length
tmp_clin2 <- tmp_clin$diagnoses %>% rbindlist() %>% data.table()
tmp_clin %>% describe()
e %>% describe()
# index the folder of folders
test <- list.files('C:/Users/thood/Documents/R/TCGA/TCGA-LAML')
# to filter by specimen case_id
tmp_biospecimen %>% select(c(case_id)) %>% arrange(case_id) %>% filter(case_id %in% 'f3c7fc84-3df8-4ff7-a378-26ec5d9e08a5')
 # matches with annotation entity_id for each zipped folder within each downloaded tcga folder
# to filter by specimen sample
tmp_biospecimen_2 %>% select(c(sample_id)) %>% arrange(sample_id) %>% filter(sample_id %in% '0a7bfd86-45c8-4959-9374-3f5166410c27')
```

TCGA File Processing

```
#code to get all the TCGA RNA-seq files
library(jsonlite)
library(R.utils)
library(dplyr)
library(Hmisc)
library(data.table)
path<- "C:/Users/thood/Documents/R/TCGA/Unzipped"
#remove any global variables
rm(list=ls())
files <- list.files(path=path, full.names=TRUE)
dataframe<-read.table(files[1])
genenames<- data.frame(dataframe[,1])
dataset<-lapply(files, read.table,colClasses = c("NULL", NA))
dataset<-cbind(genenames, dataset)
write.csv(dataset, "C:/Users/thood/Documents/R/TCGA/All594.csv")
workingfile<-read.csv("C:/Users/thood/Documents/R/TCGA/All594_60483.csv")
names_files<- read.csv("C:/Users/thood/Documents/R/TCGA/FileNames.csv", header=FALSE)
workingfile
```

DESeq2 Differential Gene Expression, Clustering, and Heatmaps

```
#libraries
library(DESeq2)
library(tibble)
##set working directory
setwd("C:/Users/thood/Documents/R/TCGA/DESeq Results EML4 vs NOn-Fused")
#remove any global variables
rm(list=ls())
#read in dataset
colData <- read.csv('clinical_recurrent.csv', row.names = 1)
dataset <- read.csv("All594_60483_Fused_recurrent_vs_normal.csv", row.names=1)
cleaned<-as.matrix(cleaned)
genes <- read.csv("gene.list.csv")
dataset_pcgenes <- dataset[dataset$dataframe...1. %in% genes$genes,]
dataset_pcgenes_trans <- t(dataset_pcgenes)
colnames(dataset_pcgenes_trans) = dataset_pcgenes_trans[1, ]
dataset_pcgenes_trans = dataset_pcgenes_trans[-1, ]
dataset_pcgenes_trans <- as.data.frame(dataset_pcgenes_trans)
dataset_pcgenes_trans <- rownames_to_column(dataset_pcgenes_trans, var="ID")
cleaned <- dataset_pcgenes_trans[dataset_pcgenes_trans$ID %in% rownames(colData),]
rownames(cleaned) = cleaned[,1]
cleaned = cleaned[,-1]
cleaned <- t(cleaned)
cleaned <- as.data.frame(cleaned)

write.csv(cleaned, "final_data_recurrent_fusions_vs_non.csv")
dataset<- read.csv("final_data_recurrent_fusions_vs_non.csv", row.names = 1)
cleaned <- as.matrix(dataset)
#check that they match
all(rownames(colData) %in% colnames(cleaned))
#put both files in the same order
cleaned <- cleaned[, rownames(colData)]
all(rownames(colData) == colnames(cleaned))
#create your DESeqData set = dds
dds <- DESeqDataSetFromMatrix(countData = cleaned, colData=colData, design = ~condition)
dds
#prefiltering
dds <- dds[rowSums(counts(dds)) > 1, ]
#creating levels for reference comparison
dds$condition <- relevel(dds$condition, ref = "Non.Fused")
dds$condition <- droplevels(dds$condition)
#differential expression analysis
#library(BiocParallel)
dds <- DESeq(dds)
res <- results(dds)
summary(res)
#look at adjusted p values
sum(res$padj < .1, na.rm=TRUE)
indices <- which(res$padj <.1, na.omit(res$padj))
genenamesPadJ10 <- rownames(res)[indices]
genenamesPadJ10
#create a csv file
```

```
write.csv(genenamesPadJ10, "genenamesPadj10_recurrent_fusions_vs_non.csv")
write.csv(res, "res_all_recurrent_fusions_vs_non.csv")
#look at .05
res05 <-results(dds, alpha = .05)
summary(res05)
sum(res05$padj < 0.05, na.rm=TRUE)
indices05 <- which(res05$padj <.05)
genenamesPadJ05 <- rownames(res05)[indices05]
genenamesPadJ05
#write a csv file
write.csv(genenamesPadJ05, "genenamesPadj05_recurrent_fusions_vs_non.csv")
write.csv(res05, "res05_recurrent_fusions_vs_non.csv")
#look at .01
res01 <-results(dds, alpha = .01)
summary(res01)
sum(res01$padj < 0.01, na.rm=TRUE)
indices01 <- which(res01$padj <.01)
genenamesPadJ01 <- rownames(res01)[indices01]
genenamesPadJ01
write.csv(genenamesPadJ01, "genenamesPadj01_recurrent_fusions_vs_non.csv")
write.csv(res01, "res01_recurrent_fusions_vs_non.csv")
#look at .001
res001 <-results(dds, alpha = .001)
summary(res001)
sum(res001$padj < 0.001, na.rm=TRUE)
indices001 <- which(res001$padj <.001)
genenamesPadJ001 <- rownames(res001)[indices001]
genenamesPadJ001
#create csv file
write.csv(genenamesPadJ001, "genenamesPadj001_recurrent_fusions_vs_non.csv")
write.csv(res001, "res001_recurrent_fusions_vs_non.csv")
#plot
plotMA(res, main="DESeq2", ylim=c(-8,8))
plotMA(res05, main="DESeq2", ylim=c(-8,8))
plotMA(res001, main="DESeq2", ylim=c(-8,8))
#pull out top adj p value genes by sorting and grabbing top 25
allres001 <- res001[indices001,]
sortbypvalue <- allres001[order(allres001$padj),]
rownames(sortbypvalue[1:25,])
#extracting normCounts
normCount <- counts(dds, normalized=TRUE)
#only get normalized counts for the genes we are interested in
normCount_all_non_genes_001 <- normCount[rownames(normCount) %in% genenamesPadJ001,]
#write normalized counts
write.csv(normCount, "recurrent_fusions_vs_non_normCount_padj001.csv")
write.csv(normCount_all_non_genes_001, "recurrent_fusions_vs_non_normCount_2219_padj001.csv")

#getting transformed values from dds
rld<- rlog(dds, blind=FALSE)
vsd <- varianceStabilizingTransformation(dds, blind=FALSE)
vsd.fast <- vst(dds, blind=FALSE)

#effects of transformation on variance
```

```
library(vsn)
notAllZero <- (rowSums(counts(dds))>0)
meanSdPlot(log2(counts(dds,normalized=TRUE)[notAllZero,] + 1))
meanSdPlot(assay(rld[notAllZero,]))
meanSdPlot(assay(vsd[notAllZero,]))
#Heatmap of count matrix
library(pheatmap)
library(grid)
select2 <- order(rowMeans(normCount_all_non_genes_001), decreasing = TRUE)
#subset the original datset for the rowmeans to get the right indices lined up
gn <- rownames(normCount_all_non_genes_001)[select2]
# defaults to log2(x+1)
nt <- normTransform(dds)
log2.norm.counts <- assay(nt)[indices001,]
datf <- as.data.frame(colData(dds)[,"condition", drop=FALSE])
map<-pheatmap(log2.norm.counts, cluster_rows=TRUE, cluster_cols = TRUE, show_rownames=FALSE,
annotation_col = datf, fontsize_col =5, main = "Log2 norm Counts")
log2.norm.counts[map$tree_col$order,]
```

ENSEMBL ID Converter

```
#get ensembl gene ids and info
library(biomaRt)
#setwd
setwd("C:/Users/thood/Documents/R/TCGA/DESeq Results Recurrent vs Normal")
#read in ensembl ids
ens <- read.csv("genenamesPadj001_recurrent_fusions_nodecimal.csv", header=TRUE)
value <- ens$gene
ensembl = useEnsembl(biomart="ensembl", dataset="hsapiens_gene_ensembl", version=79)
ids<-getBM(filters="ensembl_gene_id", attributes = c("ensembl_gene_id","entrezgene", "description",
"hgnc_symbol"),values=value, mart=ensembl)
write.csv(ids, "genenamesPadj001_recurrent_fusions_vs_normal_gene_names.csv")
```

Creating GSEA Ranked Genes List

```
#creating a rank file for GSEA
setwd("C:/Users/thood/Documents/R/TCGA")
x<-read.csv("res001_recurrent_fusions_w_gene_names_9691.csv")
x$fcSign=sign(x$log2FoldChange)
x$logP=-log10(x$padj)
x$metric=x$logP/x$fcSign

y<-x[,c("hgnc_symbol", "metric")]
y <- y[order(-y$metric),]
write.table(y,file="expression.rnk",quote=F,sep="\t",row.names=F)
```

Gene Interaction List from bioGRID

```
library(simpIntLists)
library(data.table)
setwd("C:/Users/thood/Documents/R/TCGA")
data("HumanBioGRIDInteractionOfficial")
#set up the call to find the interactions
interactions <- findInteractionList("human", "Official")
#list of genes
genes <- read.csv("genenamesPadj001_recurrent_fusions_vs_normal_gene_names_for_interactions.csv",
header=FALSE)
#get the list of interactors
test <- lapply(interactions , "[[" , "name" )
#make a data frame
df <- as.data.frame(unlist(test))
#find the indices
indices <- which(df$`unlist(test)`%in% genes$V1)
inter <-data.frame()

my_inter <- data.frame()
for (i in 1:length(indices)) {
   x <- interactions[[indices[i]]]$name
   y <- interactions[[indices[i]]]$interactors
   inter <- cbind("name"=x,"interactors"=y)
   my_inter <- rbind(my_inter, inter)
}
write.csv(my_inter, "recurrent_vs_normal_interactions_9820.csv")
```