GENE FUSIONS IN CANCER: CLASSIFICATION OF FUSION EVENTS AND REGULATION PATTERNS OF FUSION PATHWAY NEIGHBORS

by

Katelyn J. Hughes

A Thesis

Submitted to the Faculty

of

Worcester Polytechnic Institute

in partial fulfillment of the requirements for the

Degree of Masters of Science

in

Bioinformatics and Computational Biology

May 2016

APPROVED:

_____

Dr. Dmitry Korkin, Major Advisor

_____

Dr. Amity Manning, Reader

Abstract

Cancer is a leading cause of death worldwide, resulting in an estimated 1.6 million mortalities and 600,000 new cases in the US alone in 2015. Gene fusions, hybrid genes formed from two originally separated genes, are known drivers of cancer. However, gene fusions have also been found in healthy cells due to routine errors in replication. This project aims to understand the role of gene fusion in cancer. Specifically, we seek to achieve two goals. First, we would like to develop a computational method that predicts if a gene fusion event is associated with the cancer or healthy sample. Second, we would like to use this information to determine and characterize molecular mechanisms behind the gene fusion events. Recent studies have attempted to address these problems, but without explicit consideration of the fact that there are overlapping fusion events in both cancer and healthy cells. Here, we address this problem using FUsion Enriched Learning of CANcer Mutations (FUELCAN), a semi-supervised model, which classifies all overlapping fusion events as unlabeled to start. The model is trained using the known cancer and healthy samples and tested using the unlabeled dataset. Unlabeled data is classified as associated with healthy or cancer samples and the top 20 data points are put back into the training set. The process continues until all have been appropriately classified. Three datasets were analyzed from Acute Lymphoblastic Leukemia (ALL), breast cancer and colorectal cancer. We obtained similar results for both supervised and semi-supervised classification. To improve our model, we assessed the functional landscape of gene fusion events and observed that the pathway neighbors of both gene fusion partners are differentially expressed in each cancer dataset. The significant neighbors are also shown to have direct connections to cancer pathways and functions, indicating that these gene fusions are important for cancer development. Future directions include applying the acquired transcriptomic knowledge to our machine learning algorithm, counting transcription factors and kinases within the gene fusion events and their neighbors and assessing the differences between upstream and downstream effects within the pathway neighbors.

Acknowledgements:

The Bioinformatics and Computational Biology program at WPI has really helped me to grow and learn more about the field and how to apply computer science to biological problems. Thank you, first and foremost, to my advisor and mentor, Dmitry Korkin. He has really guided me through my bioinformatics education at WPI. He has been an excellent person to bounce ideas off of and voice questions and concerns about the field. He always encouraged me to do my best in every scenario. Thanks also to the rest of the Korkin Lab, for their help, encouragement and critique. Special thanks goes to Oleksandr Narykov, who gave me great feedback about my methods and helped me with the machine learning side of my project. Thank you to Amity Manning, who provided me with great comments on my thesis and methods. Thanks to the rest of the BCB department at WPI, who are comprised of wonderful role models for my research and education. Thanks to Liz Ryder and Jill Rulfs who set me up with TA funding for my two years here so I could get my MS full-time. Finally, thanks to the BBT graduate students, who critiqued my presentation. I've had a wonderful time here at WPI, and it was great being able to immerse myself in bioinformatics for the past two years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In the past decade, the discovery and optimization of next-generation sequencing (NGS) technologies has allowed us to more deeply study the intricacies of the genome. While the full potential of NGS technologies has not yet been realized and some improvements are necessary, we are able to look at our genome in a higher resolution and find both small and large chromosomal aberrations that may contribute to complex diseases such as cancer, diabetes and neurological disorders. A few of these mutations, such as copy number variants and single nucleotide polymorphisms, have been studied in great detail (Cui, 2015) and are known to be prospective targets for the treatment and cures of these diseases. Another type of these mutations, which has proven to have a direct connection with many types of cancer, are gene fusions. Gene fusions occur when two genes from distinctly separate parts of the genome, merge together to form one unit (Mertens, 2015). This event can result in a functional or non-functional gene product, depending on factors including frame shift, intact domain architecture and location of the nearest promoter (Latysheva, 2016).

1.1 CHROMOSOMAL ABNORMALITIES IN CANCER: A BRIEF HISTORY

The first known gene fusion event discovered was the Philadelphia chromosome in chronic myeloid leukemia (CML). This event resulted in the translocation of the gene ABL from chromosome 9 to chromosome 22 and merges with the BCR gene. The product is a shortened chromosome 22 and a lengthened chromosome 9 (Nowell and Hungerford, 1960). The Philadelphia Chromosome is known first mutation shown to drive cancer and this specific mutation has also been seen in other hematological cancers such as acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) (Rowley, 1973). Since the discovery of the Philadelphia chromosome, there have been many more fusion events found in both hematological and solid cancers (Mertens, 2015; Mitelman, 2007). While numerous gene fusions events have been discovered to drive cancer, many of gene fusions have been also found in healthy cells as a result of routine errors in DNA replication (Janz, 2003). These gene fusions are assumed benign but have not been studied in depth.

The purpose of this project is two-fold. Firstly, in collaboration with Oleksandr Narykov, we seek to automate the classification task of which of gene fusion events are derived identify those derived from healthy samples, which may indicate that they are benign, and those derived from cancer samples, which may indicate malignancy. We will also then take a systems approach to determine the pathway landscape of these events, and attempt to predict regulatory mechanisms for driver gene fusion events in three different cancer types: ALL, breast cancer and colorectal cancer.

1.2 MOLECULAR CHARACTERISTICS AND MECHANISMS OF GENE FUSION EVENTS

Gene fusions can occur in three ways. The most common method for developing gene fusions is through balanced chromosomal translocations. Balanced chromosomal translocations are when an arm of the chromosome swaps places with an arm of another chromosome. Depending on the location of the swap, this could cause gene fusion events in both chromosomes. The Philadelphia Chromosome (BCR-ABL) is a balanced chromosomal translocation between chromosomes 9 and 22.

Intrachromosomal deletions and inversions (Soda, 2007; Demichelis, 2007) are also able to form gene fusion events. An example of a gene fusion event formed from a deletion is the TMPRSS2-ERG fusion found in many cancers, but especially recurrent in prostate cancer. This fusion is caused by a deletion on chromosome 21 q22.2-3. This fusion is found in approximately half of all prostate cancers and contain many different junction sites causing different phenotypic effects (Perner, 2006; Clark, 2007). Inversions are formed though double strand breakage within one chromosome. These inversions can be paracentric or pericentric, targeting loci across arms or within the same arm. An example of an intrachromosomal inversion is the MLL-CALM fusion in Acute Myeloid Leukemia (Wechsler, 2003), which is an inversion of the 11^{th} chromosome at arms q14 and q23.

The gene fusion event can cause cancer in several ways. First, the gene fusion can form an oncogene, such as the BCR-ABL (Vogelstein, 2013). Second, specific domains or nucleotides in a tumor suppressor can be truncated. Third, the tumor suppressor could be moved away from its promoter and near a new promoter, giving rise to new transcriptional events.

While little is known about the differences between gene fusion events in cancer and gene fusion events in normal samples, a few bioinformatics studies have attempted to shed some light on this issue. Gene fusions frequently involve kinases, particularly tyrosine kinases (Stransky 2014; Davare, 2015), and as a result, kinase inhibitors have proven as a popular target for cancer therapy. These therapies mostly involve the kinases ALK, ETS, and RET (19). Transcription factors and histone methyltransferases have also been found at a higher probability than random chance. Complicating the matter, there have also been studies which found overlapping gene fusion events in both cancers and benign tissue. For example, in one study, the gene fusion VTI1A-TCF7L2 has been found in 42% of colon cancer but also in 29% of normal colon tissue and 25% of normal tissues from other organs (Nome, 2014). Normal gene fusions are hypothesized to be a way to increase proteomic diversity (Parra, 2006; Casado-Vela, 2013), but more research needs to be done to identify protein formation from gene fusion events on a large scale to determine other underlying causes and effects.

1.3 CURRENT GENE FUSION DETECTION TOOLS

1.3.1 TOOLS TO DETECT GENE FUSION EVENTS IN RNA-SEQ

Our increasing ability for high-throughput sequencing spawned a need for new computational methods to detect specific mutations in our sequences. Currently, there are numerous tools (Iyer, *et al*., 2011, Ge, *et al*., 2011; Kim, *et al.*, 2011; McPherson, *et al*., 2011; Abate, *et al*., 2012; Francis, *et al.*, 2012; Beccuti, *et* al., 2014; Fernandez-Cuesta, *et al*., 2015; Davidson *et al*., 2015) to detect specific gene fusion events directly from both paired-end and single read RNA-seq and Whole Genome Sequencing (WGS). These software methods are useful for detecting known gene fusion events as well as new gene fusion events through looking at the breakpoint junctions.

Most of these tools can be classified into three different types, as explained in Cararra, *et al.*, 2013. Whole paired-end based tools align the paired-end reads to a reference genome and use conflicting alignments to generate a set of fusion events. These events are then filtered based on a series of characteristics of real gene fusion events. Some of the tools that use this are deFuse

(McPherson, 2011) and FusionHunter (Li, 2011). In whole paired-end plus fragmentation methods, reads are also aligned to the genome but conflicting alignments are used to generate a new reference genome. In addition, the unaligned reads are also fragmented and realigned to the new fusion reference genome. TopHat-fusion (Kim, 2011), ChimeraScan (Iyer, 2011) and Bellerophontes (Abate, 2012) are methods that fall in this category. The final category is direct fragmentation based tools, which fragments the reads first and then aligns those fragments to a reference. FusionMap (Ge, 2011) and MapSplice (Wang, 2010) are examples of direct fragmentation fusion detection tools.

Independent studies that have analyzed these methods in depth for time complexity, accuracy and the number of false positives have given insight into the variance of gene fusion detection methods (Wang, 2013; Liu, 2016; Cararra, 2013). Unfortunately, there is very little overlap between software and there is a tendency for false positives in many of these programs. One method, FusionMap, balances specificity in negative datasets and sensitivity in positive datasets analyzed (Carrara, *et al*., 2013). Despite the number of methods already developed, the best way to determine accurate gene fusion events with our current NGS technology is still in the process of being researched. The emergence of long-read sequencing could assist these methods in more accurately detecting fusion events, although many of these methods have not been tested on reads longer than 100 bp. In our analysis, we utilize the features produced by FusionMap, but may incorporate more software methods in the future.

1.3.2 TOOLS TO DETECT MALIGNANT GENE FUSION EVENTS IN CANCER

Two methods currently exist to assess the oncogenic potential of the gene fusion event. Oncofuse is a supervised naïve bayes classifier, which utilizes intact domain information to determine a fusion events relationship with cancer (Shugay, *et al*., 2013). Oncofuse uses a very broad classifier, which contains gene fusion information from all cancers in one model. Accounting for the observation that gene fusion events are tissue-specific, Pegasus relies on a more advanced supervised learning method, boosted decision trees, to predict oncogenic potential in glioblastoma multiforme and

anaplastic large cell lymphoma (Abate, *et al*., 2014). Pegasus, uses a more specific model, contributing the idea that gene fusion events are cancer-specific. Both of these methods exhibit high-accuracy in fusion events through feature-based supervised learning.

While Oncofuse and Pegasus work well for their specific cases, an improved model is required. Preliminary analysis has discovered that for each cancer in this analysis (acute lymphoblastic leukemia, breast cancer and colorectal cancer), there is over one third of the fusion events which overlap between cancer and healthy (Figure 1). Previous studies have incorporated the data without reclassifying it as belonging to both categories or disposed of the data completely. To use and correctly classify this data, we employed a semi-supervised method which finds these overlapping fusion events, classifies them as unlabeled at initiation and attempts to classify each event based on the characteristics of similar events.

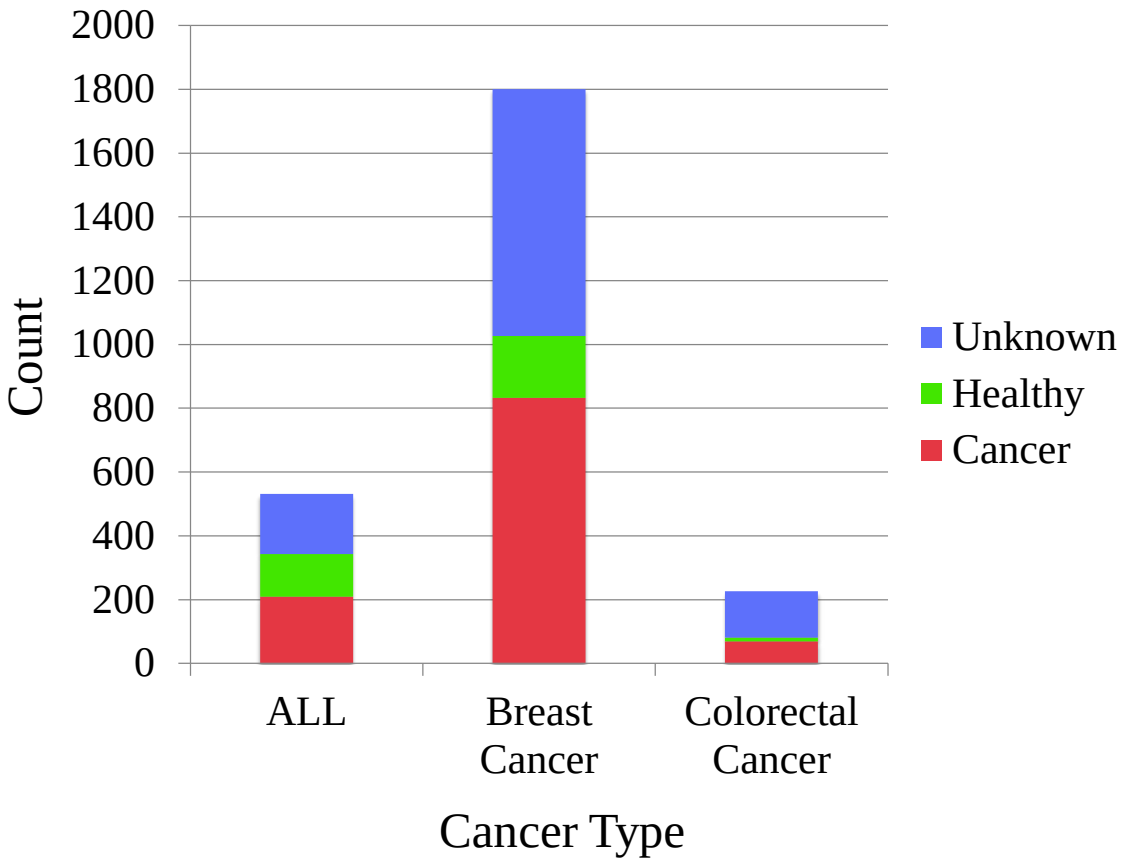

Figure 1: Classification count of gene fusion types for each cancer studied.

1.4 MACHINE LEARNING FOR CLASSIFICATION

1.4.1 COMPARISON BETWEEN SUPERVISED AND SEMI-SUPERVISED METHODS

Both Pegasus and Oncofuse use a method called supervised machine learning, which trains the algorithm to learn whether a gene fusion event comes from cancer or healthy samples using already known examples from cancer and healthy samples. The mathematical model is then tested blindly on known examples, and the correctness of the algorithm is tested by comparing the number of guesses by the algorithm in each group to the actual number of examples in each group. Supervised learning algorithms are usually tested at 60% training set and a 40% testing set. For more information about the specific methods, equations and methods to assess an algorithms correctness, see the methods section.

Instead, for our method we use semi-supervised learning. Our algorithm utilizes both the unclassified (fusion events that occur in both healthy and cancer) and the unique events in cancer and healthy samples for training the model. To begin with, all data points are classified as unlabeled, cancer or healthy. The algorithm then tries to classify all of the unlabeled points, placing the top 20 best predictions back into the training set. The other predictions are returned to the unlabeled class and the process is repeated until all of the points are labeled.

1.5 GENE EXPRESSION REGULATION IN GENE FUSION PATHWAY NEIGHBORS

Other more heavily studied somatic mutations (e.g. SNV and CNV) are known to affect gene expression in the cell (Haraksingh, 2013). Most of the gene expression variation that affects SNVs is associated with the frame-shift which happens if the mutation is located in the coding region. The connection between gene expression and copy number variants is more obvious, as it directly contributes to gene dosage in cells. However, due to our current lack of understanding of gene fusion events, this phenomenon has not been studied in gene fusion events.

Through the analysis of the machine learning methods, we determined that some cancers performed better in our algorithm than others and as a result, we wanted to further explore why. The double strand breakage and chromosomal abnormality that results from can cause dysfunction within

the cell if circumstances are correct. In this case, we would need an intact, appropriate promoter as well as a start codon, a stop codon and no apparent frameshift.

A biological pathway is a series of molecules which produce a particular reaction within the cell. Members of these pathways can be affected by numerous pre and post-transcriptional and translational mechanisms such as gene dosage, gene expression, alternative splicing and post-translational modification to name a few. One way to study the effects of the gene fusions on the pathway is to measure the gene expression of its neighbors. It is predicted there will be a ripple effect of these gene fusion events occurring in the genome, which will cause the genes surrounding the gene fusion event partners in a pathway to be differentially expressed. This ripple effect is predicted to present itself as differential expression of genes associated with cancer pathways of unique gene fusion pathway neighbors in both cancer and healthy samples.

2 HYPOTHESES AND GOALS

The main goal of this project is to understand the role of gene fusion events in cancer and provide insights into the molecular mechanisms behind cancer-associated gene fusions. We achieve this goals through two specific aims. **First**, we develop a semi-supervised machine learning method for the classification of healthy and malignant gene fusion events. It is expected that as a result of our semi-supervised method leveraging the unlabeled gene fusion events as more related to the cancer or healthy sample, our model to classify the already known gene fusion events will improve.

Second, we explore the large-scale effects of gene fusions through pathway analysis genes. We hypothesize that we will see some overlapping pathways and genes. We then explore the unique significant genes who are pathway neighbors of gene fusion. We expect to derive much of the landscape of gene expression patterns for each of the gene fusion partners' pathway neighbors. our second hypothesis is that we will find some genes within pathway neighbors that are differentially expressed, indicating that changes on the structural level also have an impact on the transcriptomic level, due to frame shift changes and partially intact domains. We also expect that many of these genes

will be involved with cancer pathways.

3. METHODS

3.1 NEXT GENERATION SEQUENCING ACQUISITION AND FUSION PREPROCESSING

Paired end RNA-seq samples were acquired from NCBI's Sequence Read Archive (SRA). Adapter sequences and low quality reads were trimmed using Trimmomatic Version 0.36 (Bolger, 2014). Reads were omitted if they fell below 20 base pairs and nucleotides below a quality score (PHRED33) of 3 were removed. Trimmomatic scans the reads with a 4 base wide sliding window and cuts when the average quality per base drops below a PHRED33 quality score of 15. To detect fusion events, all trimmed FASTQ files were used as input for FusionMap (Ge, 2011). Human genome build 37.3 was used to map fusion events to the genome. Features detected from FusionMap were used in both supervised and semi-supervised learning algorithms. Fusion events were screened for previous discovery as a recurrent gene fusion event in cancer using Chimera DB tissue search.

3.2 DATASETS

To explore the breadth of cancers known to harbor gene fusions, Illumina paired end RNA-seq datasets from acute lymphoblastic leukemia (ALL), breast cancer (BC) and colorectal cancer (CRC) were obtained and analyzed using the above methodology. Samples were obtained from NCBI SRA for subsequent analyses.

ALL was used as a representative of the hematological cancers, for which gene fusion events have been well-characterized. As described in Almamun, *et al.*, 2015, the dataset was comprised of 20 pre-B ALL patient samples and 10 pre-BI and pre-BII samples isolated from human cord blood. Samples were obtained from NCBI SRA dataset SRP058414. A total of 534 gene fusion events are found, 209 occurring solely in cancer, 134 occurring solely in healthy samples and 189 of unknown origin as they were found in both cancer and healthy samples.

Colorectal cancer was used as a representative of a smaller matched study. As explained in Kim, *et al.*, 2014, the dataset was comprised of 18 primary CRC tumors and 18 matched normal colon

samples. Samples were obtained from NCBI SRA dataset SRP029880. A total of 226 gene fusion events were found with 69 occurring uniquely in cancer, 12 occurring uniquely in healthy samples and 145 of unknown origin.

Breast cancer was used as a representation of a large scale matched study. As explained in Varley, *et al.*, 2014, the dataset was comprised of 28 breast cancer cell lines, 42 triple negative breast cancer (TNBC) primary tumors, 42 estrogen receptor positive (ER+) breast cancer primary tumors, and 56 healthy breast tissue samples. Triple negative breast cancer lacks expression of the estrogen receptor, progesterone receptor and HER2/neu. ER+ samples have the estrogen receptor expressed. Within these samples, there are 26 matched ER+ samples and 17 matched TNBC samples. Except where noted, most analyses are performed just using the matched samples. Samples were obtained from SRA dataset SRP042620. For the matched dataset, a total of 1773 gene fusion events were found with 833 occurring solely in cancer, 194 occurring in healthy samples and 744 of unknown origin occurring in both cancer and healthy samples.

3.3 FEATURES REPRESENTING GENE FUSION EVENTS

Features (Table 1) generated from FusionMap were used to detect structural characteristics of gene fusion events. To account for the importance of the relationship between the fusion partners, our group developed a gene ontology classification system for each fusion partner separately. Each gene ontology section (Biological Process, Cellular Component and Molecular Function) is a series of graphs, each with a node and a vector connecting the node to its ancestors and descendants. We found all descendants of the three sections up to three descendants away from the top node (Biological Process, Cellular Component, and Molecular Function) (Figure 2). Due to the large number of genes in the top node, the top node was not included in this analysis. Due to the large number of GO terms and high level of redundancy amongst the terms, any GO term with a distance of greater than 3 from the root node was omitted from the feature set. For each given gene fusion event, involving a pair of genes, and a given GO-term the numerical feature corresponding to that term was 0 if neither of the genes

were associated with this GO term, 1 if only one of the genes was associated with the GO term, and 2 if both genes were associated with the GO term. The resulting feature set includes 36,032 features, a large number given a relatively small training set, which could contribute to severe underfitting of our model.

Since we have a large feature set, it is proposed that we could use dimensionality reduction to decrease the number of features. Gene ontology terms have an extremely repetitive vocabulary and as a result, our model could be improved by some techniques from natural language processing, most specifically topic modeling. From these methods, we were able to cut down the number of features from 36032 features for each dataset to around 50 features for each dataset.

3.4 SUPERVISED LEARNING METHODS

As a baseline to explore our novel semi-supervised algorithm, supervised learning classification algorithms Naive Bayes, SVM and Random Forest were used. For this set of experiments, ambiguous gene fusion events, belonging to both cancer and healthy samples, were not used.

Naive Bayes bases decisions on the idea that given a particular classification, what is the probability that a particular attribute can occur? Naïve Bayes was selected because this is the model that Oncofuse uses in for its datasets. The classifier (Mitchell, 1998) assumes conditional independence of X containing n attributes, which are conditionally independent of one another given Y. When a new instance of X is introduced, Naive Bayes will calculate the probability that Y will take on any value given the attribute values of the new X and the distributions of $P(Y)$, and $P(X_i|Y)$.

Support vector machines (Cortes, Vapnik, 1995) are a classification algorithm which first determines the points that lie closest to the selected decision boundary. These points may be ambiguous in terms of whether they belong to the positive or negative datasets. These points are denoted as the support vectors, and the maximum margin from the distance between the support vectors and the decision boundary is determined. This is established through rotation of the entire dataset. The type of support vector machine used here is called radial basis function, which works well for non-linear

datasets. The classical linear SVM solves the following optimization problem:

$$\begin{cases} \frac{1}{2}\| \quad \omega\|^2 + C\sum_{n=1}^{N} \xi_n \rightarrow \min\limits_{\omega,b,\xi}, \\ t_n(\omega^T x_n + b) \geq 1 - \xi_n, \\ \xi_n \geq 0 \end{cases}$$

where x_n is a feature vector for n-th data point in our dataset, $t_n \in \{-1, +1\}$ is it's label, ω is the weights vector, b is a bias, ξ_n is a cost for misclassifying sample which depends upon distance from the margin and C is regularization parameter.

Random forests (Breiman, 2001) uses a combination of decision trees and bagging in order to determine the class of a particular vector. For a specific number of trees, in our case 100, a number of features, in our case 3, are selected at random. With these three features, a decision tree is created in order to classify the sample as belonging to cancer or normal samples. The best predictors for each node is selected from random subsets. Class is predicted based on majority vote of the resulting trees. As a result of their random nature, random forest models work great for heterogeneous datasets.

3.5 SEMI-SUPERVISED LEARNING METHODS

Semi-supervised learning (Ratsaby and Venkatesh, 1995) is a class of algorithms for which you only know some of the classifications of the samples before you begin the algorithm. Since random forest was the highest performing method, it was applied to our semi-supervised method to determine the correct classification for each unlabeled point. The top 20 definitive points were kept and placed back into the training set. The process was repeated until all points were classified. Finally, the testing set here is comprised of only healthy and cancer fusion events that are not overlapping between the two groups.

3.6 MACHINE LEARNING ALGORITHM ASSESSMENT

Our algorithm will likely classify some fusion events right and some fusion events wrong. To determine the best algorithm for each dataset, we measure the true positives, true negatives, false positives and false negatives. Table 2 illustrates these points. When a positive data point is correctly

predicted as a positive data point, it is referred to as a true positive. When a negative data point is correctly predicted as a negative data point, it is referred to as a true negative. When a positive data point is incorrectly predicted as a negative data point, it is referred to as a false negative. When a negative data point is incorrectly predicted as a positive data point it is referred to as a false positive.

Through these measurements, we can now determine various metrics of our algorithm. Accuracy is the total percentage the algorithm gets correct. It is normalized by the total size of the dataset and includes all positives and all negatives.

$$A = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$$

Precision is the total percentage of the true positives when taking into account all of the positive values. For our situation, this translates into the number of times the algorithm guessed that the gene fusion event was cancer and it was correct, compared to the total amount of times that it guessed cancer.

$$P = \frac{T_P}{T_P + F_P}$$

Recall is the total percentage of the true positive samples that were selected by the algorithm compared to the total number of actually positive samples. For our situation, this translates to the number of times our algorithm was able to extract the positive values out of all of the positive values.

$$R = \frac{T_P}{T_P + F_N}$$

Finally, we use the F-score as a combined metric to assess the algorithm's efficiency. It combines the scores for both precision and recall.

$$F = 2\,\frac{P\ x\ R}{P + R}$$

3.7 RNA-SEQUENCING ANALYSIS

Gene expression analysis was performed using the Tuxedo Suite. Tophat (Trapnell, 2009) was used to align reads to the genome, Cufflinks (Trapnell, 2010; Trapnell, 2013) was used to quantify the total number of Fragments per Kilobase of transcript per Million mapped reads (FPKM) which indicates the expression values for all of the groups. This is a normalized measure which takes normalizes the fragment found and to the length of the read. Tophat accepts RNA-seq reads data and then aligns the reads to a known genome. This will enable us to determine annotation for the already known genes. The maximum, average and median FKPM were taken for all groups, as well as a p-value to assess significance of the expression level for particular genes.

3.7.1 PATHWAY ANALYSIS

Reactome (Fabregat, 2016; Milacic, 2016) pathway neighbors of either fusion partner was determined using Pathway Commons (Cerami, 2011) neighborhood function. The Reactome database was selected over other commonly used pathway databases (KEGG, PANTHER, etc.) because of its in-depth biochemical pathways. Initial analysis showed the glycolysis pathway may be affected and we were interested in more specific steps of glycolysis than KEGG could provide. Neighbors were selected at a distance of two or less nodes upstream or downstream from the fusion partner. Delta FKPM was detected through subtracting the averaged healthy from the averaged normal FPKM. The most significant neighbors were detected by calculating the mean and standard deviation for the healthy neighbors and then applying it to the cancer neighbors to find if there are any neighbors fall outside of the normal distribution of the healthy neighbors. Cancer fusion neighbors with healthy-normalized z scores greater than 3σ were explored further through literature review and also compared to the normal fusion neighbors. Distribution of both healthy and cancer fusion neighbors were determined and enrichment was calculated using the following formula. Overlapping genes between cancers were identified to assess potential drug targets. Unique genes within cancers and between cancer and healthy from the same tissue were also assessed.

Enrichment of genes was determined via the software program Enrichr (Chen, 2013). Enrichr is a pathway analysis tool which uses Fisher's exact test for detecting which pathways are enriched in a phenotype of choice. We chose this algorithm because it tends to not bias towards large pathways and it allows us to specifically pick the commonly known pathway database of our choice. In our case, we used Reactome. A p-value is determined based on Fisher's exact test, then a z-score is calculated based on intuition. Finally, the two methods are put together to create one combined score:

$$c = \log(\mathrm{p})\,\mathrm{z}$$

In this equation, c is the combined score, p is the p-value which is determined by Fisher's exact test, and z is the deviation of the pathway from the expected rank as developed through intuition.

Table 1: FusionMap features used for machine learning analysis

FusionMap Features	**Original Type?**	**Definition**
Seed Count	Numerical	Unique fusion junction cut sites
Rescued Count	Numerical	Number of reads for a specified fusion.
Position1 (5')	Numerical	What position is it at in the genome
Position2 (3')	Numerical	
Exon Number1 (5')	Numerical	What is the exon number?
Exon Number2 (3')	Numerical	
Strand	Categorical	Does it come from a positive or negative strand?
Splice Pattern Class	Categorical	What is the specific pattern at the fusion?
Frame Shift Class	Categorical	Is there a frameshift mutation?
Breakpoint distance	Numerical	Distance between original cut sites.
On Exon Boundary	Categorical	Does this lie on an exon boundary?
Chromosome number 1 (5')	Numerical	Original chromosome number for each gene.
Chromosome number 2 (3')	Numerical	
Label	Categorical (Target)	Does this event belong to cancer or healthy?

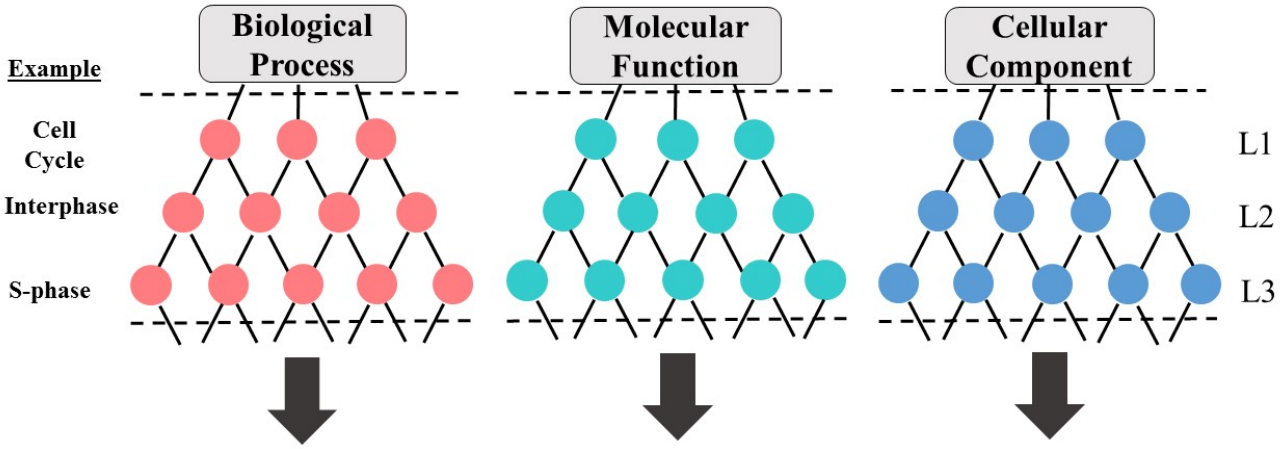

Figure 2: Algorithm of calculating Gene Ontology (GO) features defined. Each of the three sub-categories of Gene Ontology was treated as a multi-layer graph. The top three GO layers were used as our functional features.

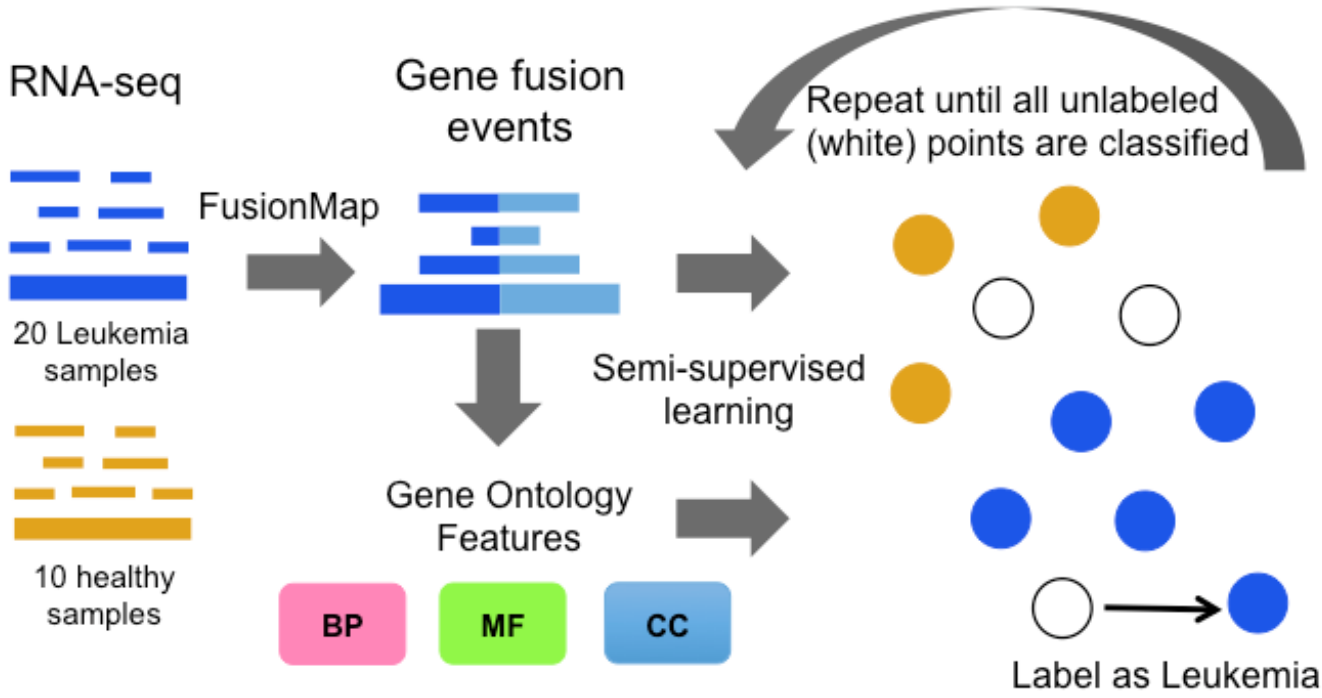

Figure 3: Example of procedure of self-iterative semi-supervised learning, represented in Leukemia dataset. Leukemia samples are trimmed and ran through FusionMap software and then separated into only present in healthy, only present in cancer or unlabeled (present in both). The unlabeled points are classified by random forest. Then, the top 20 most unambiguous points are reintegrated into the training set and the rest are returned to the unlabeled class. The process is repeated until all of the unlabeled points are labeled as either cancer or healthy.

Table 2: Machine learning assessment for semi-supervised and supervised learning

		Actual	
		Cancer	**Healthy**
predicted	**Cancer**	True Positive	False Positive
	Healthy	False Negative	True Negative

4. RESULTS

4.1 PRELIMINARY MACHINE LEARNING ANALYSIS OF ACUTE LYMPHOBLASTIC LEUKEMIA

To test how well each supervised algorithm performed, we subjected our supervised Naïve Bayes, Support Vector Machines (SVM) and random forest on our acute lymphoblastic leukemia dataset, before applying feature selection. We also applied our semi-supervised algorithm to this analysis. From these results, it was determined that random forest performed the best out of all of the supervised algorithms, since it had the best recall (77.78%), accuracy (84.06%) and F-score (79.25%). Naïve Bayes performed the worst in accuracy, precision and F-score. Support vector machines had the best precision (97.05%), but also had the worst recall (61.11%). Since our best algorithm for this dataset was random forest, the self-iterating semi-supervised random forest was then applied to this dataset. Using this algorithm gave us an increase of three to six percentage points in accuracy. It also the highest accuracy (87.68%) as compared with the supervised models, as well as the highest F-score (82.47%). (Table 3**)**

Table 3: Supervised and semi-supervised learning of Acute Lymphoblastic Leukemia, pre-feature selection. Values in red indicate worst performance. Values in green indicate best performance

Method	Accuracy	Precision	Recall	F-score
Naïve Bayes	81.1%	78%	72.2%	75%
SVM	84.0%	97.0%	61.1%	75%
Random Forest	84.0%	80.7%	77.7%	79.2%
Semi-supervised	**87.6%**	**93.0%**	**74.0%**	**82.4%**

4.2 MACHINE LEARNING ANALYSIS OF CANCER DATASETS WITH FEATURE SELECTION

Since our best working algorithms for ALL were our semi-supervised and supervised random forest algorithms, it was decided to apply those algorithms to matched breast cancer dataset. Using a matched dataset will allow us to determine if our algorithm works on matched patients. When breast cancers were separated into ER+ (estrogen receptor expressed) and triple negative breast cancer (lacking estrogen receptor, progesterone receptor and HER2/neu expression), we had high accuracy and F-measure using only supervised random forest (Table 4). This is before using our topic modeling feature selection at 37 total Gene Ontology topics along with our FusionMap features. All future machine learning protocols will use breast cancer phenotypes separately. Figure 4 summarizes accuracy and F-score for random forest without feature selection amongst all cancers. Not including feature selection, ER+ had an accuracy of 89.74% and an F-score 82.61%. TNBC had an accuracy of 83.33% and an F-score of 83.23%. Leukemia had an accuracy of 85.4% and an F-score of 81.93%. Colorectal cancer had an accuracy of 84%. Colorectal cancer had a very low gene fusion normal count (only 12 gene fusions were detected as being unique in normal colon tissue), which is why we were not able to obtain an F-score for colorectal cancer.

While it may be more accurate to have many features, it doesn't make sense for biological clarity to have so many redundant features. Figure 5 summarizes supervised random forest F-score and accuracy for all cancers with 37 total Gene Ontology features derived from topic modeling feature selection and FusionMap features. Our overall performance for supervised random forest was consistent across cancers. Including feature selection, ER+ had an accuracy 80.3% and an F-score of 61.01%. TNBC had an accuracy of 83.33% and an F-score of 83.23%. Leukemia had an accuracy of 86.4% and an F-score of 82.5%. Colorectal cancer had an accuracy of 84%. Colorectal cancer had a very low gene fusion normal count (only 12 gene fusions were detected as being unique in normal colon tissue), which is why we were not able to obtain an F-score for colorectal cancer. In general, this

model didn't work as well as the model without feature selection, but it has the benefit of only being around 50 features as opposed to over 30,000. This is beneficial for two reasons. First, it will speed up our algorithm greatly. Second, it will make this more biologically relevant, because we are combining all similar features into one.

Next, we tried semi-supervised learning using the above methods. Figure 6 shows the F-score and accuracy performance of our semi-supervised method using feature selection through topic modeling. Including feature selection, ER+ had an accuracy 83.7% and an F-score of 68.85%. TNBC had an accuracy of 83.33% and an F-score of 83.23%. Leukemia had an accuracy of 83.4% and an F-score of 79.01%. Colorectal cancer had an accuracy of 84%. Colorectal cancer had a very low gene fusion normal count (only 12 gene fusions were detected as being unique in normal colon tissue), which is why we were not able to obtain an F-score for colorectal cancer. Overall, our semi-supervised model performed about equally with the supervised model, indicating that we may need some different features in order to improve the score.

Table 4: Supervised random forest for breast cancer subsets

Method	Accuracy	Precision	Recall	F-score
Random Forest TNBC	86.6%	86.6%	86.6%	86.4%
Random Forest ER+	89.7%	95%	73.0 %	82.2%

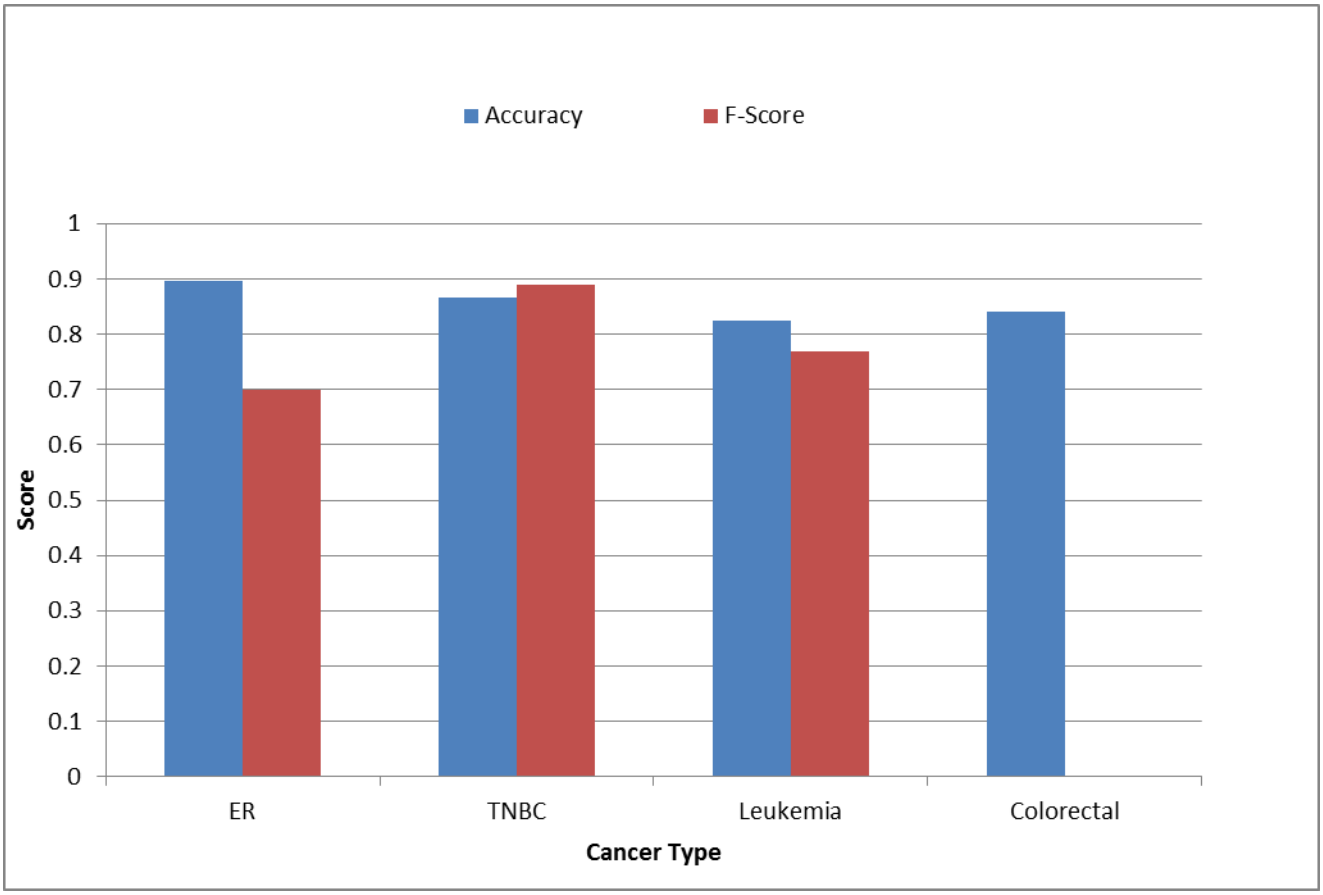

Figure 4: Supervised random forest for all cancers studied with no topic modeling included. Blue indicates accuracy and red indicates F-score

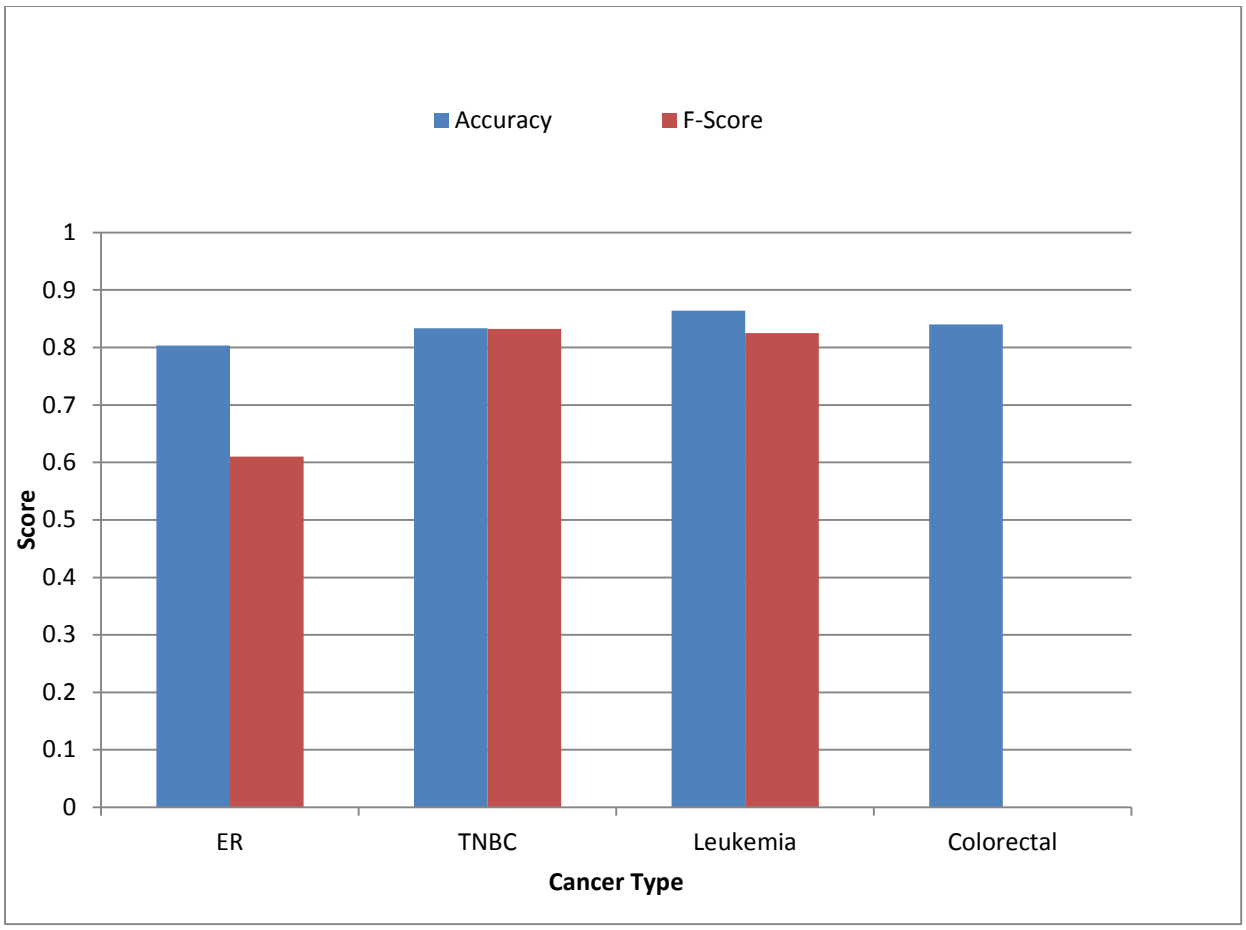

Figure 5: Supervised random forest for all cancers studied with LDA topic modeling included. Blue indicates accuracy and red indicates F-score

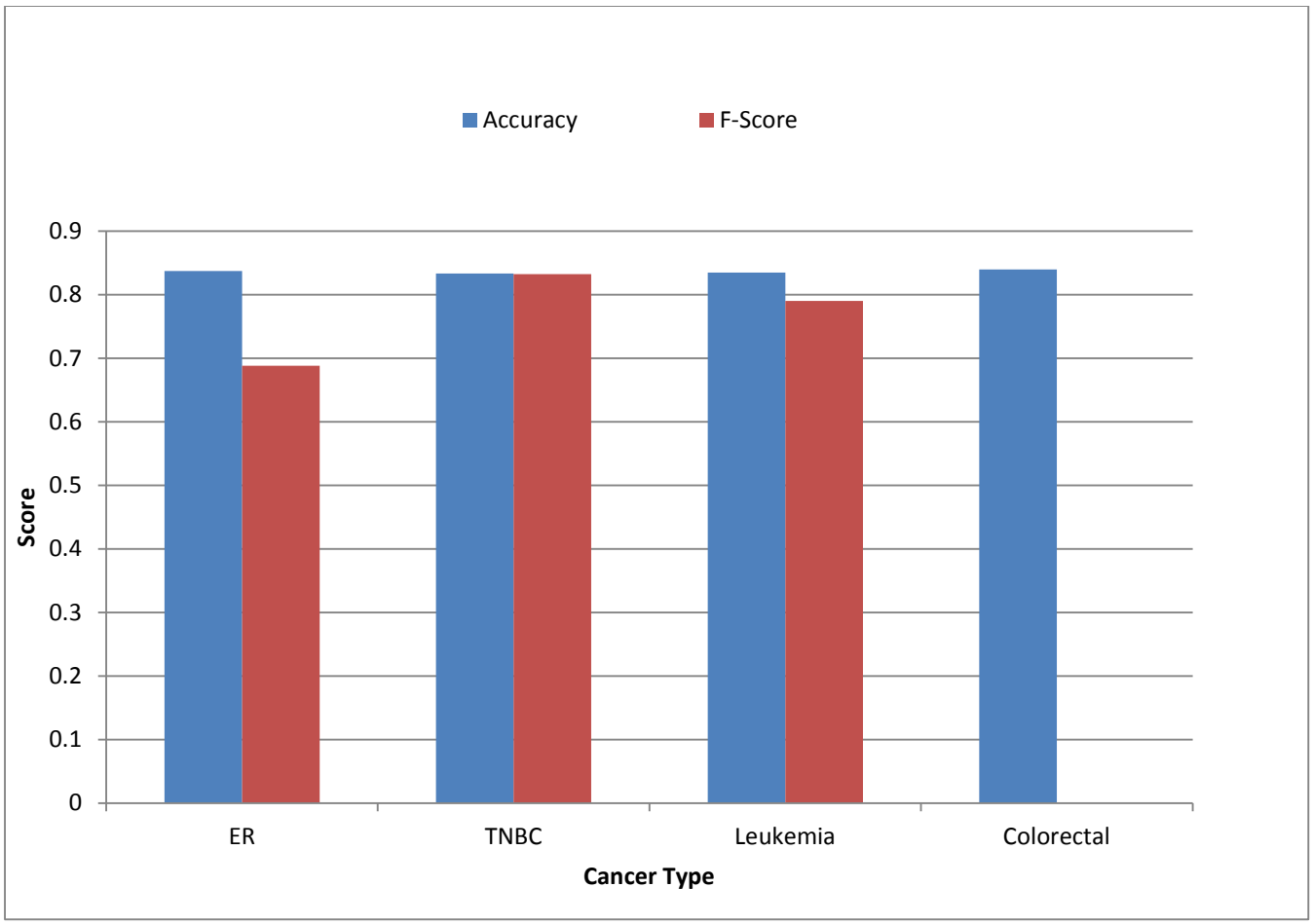

Figure 6: Semi-supervised random forest for all cancers studied with LDA topic modeling included.

Blue indicates accuracy and red indicates F-score

4.3 OVERALL GENE FUSION ANALYSIS

To begin to understand the landscape of the gene fusion events detected using FusionMap in the three cancers selected, an overall gene fusion pathway analysis and literature review was performed for all of the gene fusions retrieved. This will assure us that these gene fusion events are not artifacts of next-generation sequencing errors or false positives. Out of all of the gene fusion events in the three datasets, only HLA-A-HLA-B (Kimura, 2006) from colorectal cancer, POLA2-CDC42EP2 (Kang, 2016), and fusions involving FGF1-3 (Kumar-Sinha, 2015), have been previously identified as recurrent gene fusion events as reported by ChimeraDB. However, we also found genes that are family members with genes known as recurrent gene fusion events such as ETV2, which is in the ETS transcription factor family, related with ETV6, which is involved in a known gene fusion event in ALL (TEL-AML1). The other gene fusion events could either be false positives or novel gene fusion events. More analysis would be needed with another gene fusion software program to determine whether or not there are false positives within these datasets and determining overlap between the software programs.

4.4. GENE FUSION PATHWAY ANALYSIS

Pathway analysis of all of the gene fusion events were determined via Enrichr (Chen, 2013). Enrichr was used as an alternative to Gene Set Enrichment Analysis (Subramanian, 2005) and the multitude of other enrichment software programs for a few reasons. Firstly, both gene set enrichment analysis and other enrichment programs have a tendency to be biased towards large gene sets. Enrichr uses a more advanced algorithm to prevent this. The algorithm first computes the p-value from Fisher's exact test, which indicates the likelihood that detecting those focus genes in the pathway is not due to random chance (p-value). This method is used among many of the enrichment software programs. Additionally, it computes the deviation of the expected rank, which was developed based on intuition and through testing on multiple datasets, compared with the actual rank (z-score).

Secondly, Enrichr uses thirty-five well-known databases as its gene set libraries. For comparison purposes, GSEA uses a set of seven gene set libraries that are primarily made through

manual curation. As a result, in Enrichr the user can select which pathway database to use in order to obtain consistency with the rest of their experiment. For this instance, Reactome pathways were used in order to stay consistent with later studies.

All of our gene fusion events were used as queries for the Enrichr software. In order to view the differences between cancer and normal tissue or cells in each cancer, gene fusion events were divided based on cancer type and also whether the gene fusion event was associated in our dataset with normal or cancer. Unclassified gene fusion events were omitted from this analysis.

Figures 7-12 show the results from this study. As expected, due to the tissue-specific nature of gene fusion events, enriched pathways rarely overlap between the genes involved in gene fusions across cancers and between healthy and malignant tissue or cells. Tables 5-10 show specific genes in each dataset associated with each pathway. From these tables, it has been observed that there are very few overlapping gene fusion associated genes that contribute to the enrichment of these pathways. This can be said for genes that are overlapping between cancer fusions and healthy fusions, as well as genes that span cancers. This may fuel the argument that gene fusions are tissue specific. Many of the pathways that are enriched are similar within the group, as seen with the ALL cancer fusion partners.

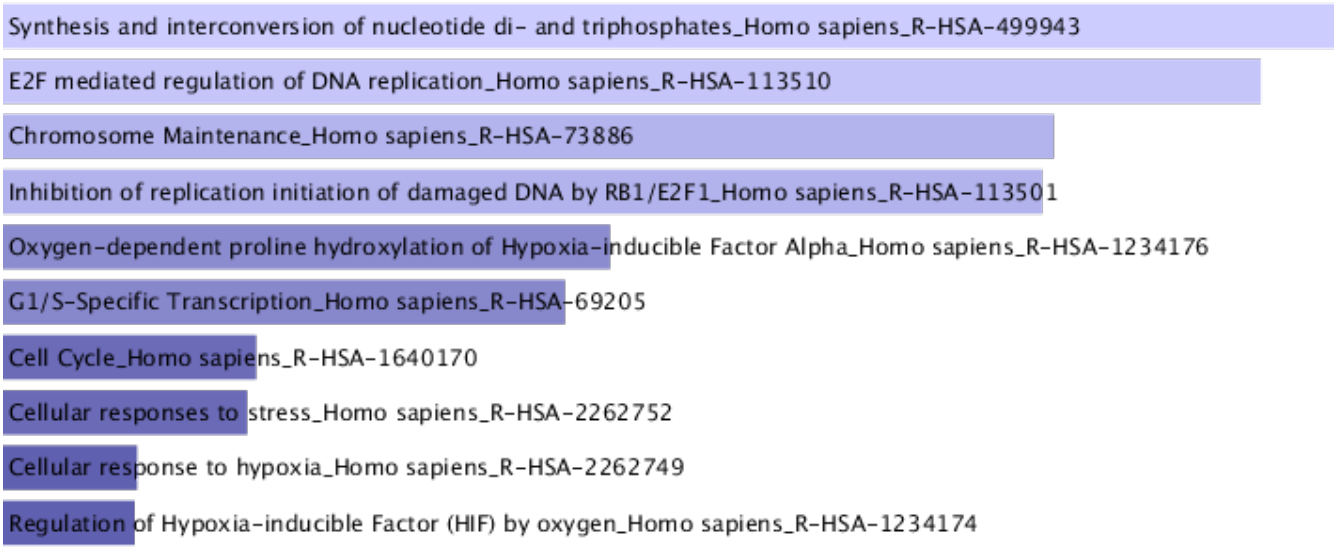

Figure 7: Enriched pathways for genes associated with cancer gene fusion events in ALL as determined by Enrichr**.** Length of bar directly correlates to combined score as generated by Enrichr.

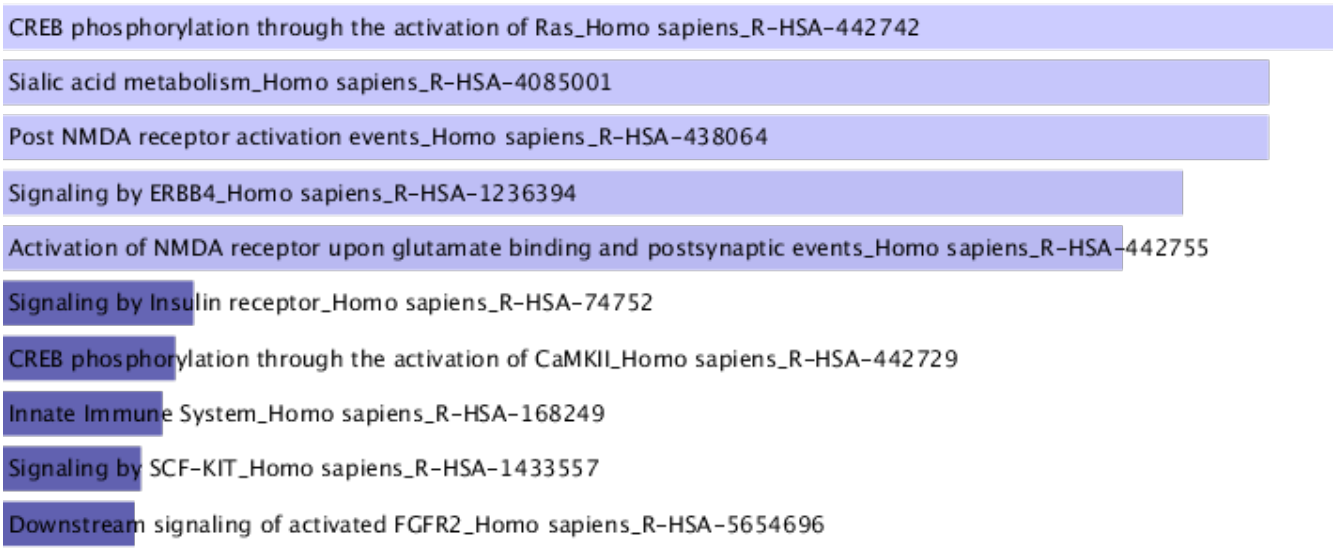

Figure 8: Enriched pathways associated with healthy gene fusion events in ALL as determined by Enrichr ranked by enrichment score. Length of bar directly correlates to combined score as generated by Enrichr.

Table 5: Details of enriched pathways in gene fusions represented in acute lymphoblastic leukemia. P-value represents the p-value as determined by Fisher's exact test. Z-score is the z-score determined by the enrichr algorithm which is based on intuition and combined is the combined score between the p-value and the z-score.

Term	P-value	Z-score	Combined	Genes
Synthesis and interconversion of nucleotide di- and triphosphates	0.001	-2.100	2.609	RRM2;NME4;AK7
E2F mediated regulation of DNA replication_Homo sapiens	0.002	-2.109	2.518	POLA2;RRM2; E2F1
Chromosome Maintenance_Homo sapiens	0.003	-2.061	2.262	POLA2;APITD1; SMARCA5;TERF2
Inhibition of replication initiation of damaged DNA by RB1/E2F1	0.004	-2.048	2.247	POLA2;E2F1
Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha	0.008	-2.160	1.711	UBE2D2;HIF1A
G1/S-Specific Transcription_Homo sapiens	0.007	-2.090	1.655	RRM2;E2F1
Cell Cycle	0.035	-2.344	1.272	POLA2;APITD1; RRM2;STAG; DMC1;SMARCA5; E2F1;TERF2
Cellular responses to stress	0.012	-2.319	1.261	PRKAA1;UBE2D2; E2F1;CHMP4A; BMI1;TERF2;HIF1 A
Regulation of Hypoxia-inducible Factor (HIF) by oxygen	0.015	-2.080	1.131	UBE2D2;HIF1A
Cellular response to hypoxia	0.015	-2.050	1.114	UBE2D2;HIF1A

Table 6: Details of enriched pathways in gene fusions represented in healthy blood. P-value represents the p-value as determined by Fisher's exact test. Z-score is the z-score determined by the enrichr algorithm which is based on intuition and combined is the combined score between the p-value and the z-score.

Term	P-value	Z-score	Combined	Genes
CREB phosphorylation through the activation of Ras	0.001	-2.136	3.985	PDPK1;CAMK2A; CAMK2G
Sialic acid metabolism	0.001	-2.099	3.915	NEU3; ST6GALNAC4; ST6GALNAC6
Post NMDA receptor activation events	0.001	-2.098	3.914	PDPK1;CAMK2A; CAMK2G
Signaling by ERBB4	0.003	-2.542	3.821	NR4A1;NCSTN; ITCH;PDPK1; CAMK2A;IL2RG; CAMK2G
Activation of NMDA receptor upon glutamate binding and postsynaptic events	0.002	-2.018	3.756	PDPK1;CAMK2A; CAMK2G
Signaling by Insulin receptor	0.037	-2.287	2.747	PDPK1;CAMK2A; IL2RG;ATP6V0C; CAMK2G
CREB phosphorylation through the activation of CaMKII	0.004	-1.833	2.727	CAMK2A; CAMK2G
Innate Immune System	0.019	-2.259	2.713	NR4A1;ITCH; PDPK1;ARPC1B; CAMK2A;ARPC4; NLRP1;IL2RG; MAP3K14; CAMK2G
Signaling by SCF-KIT	0.043	-2.240	2.690	NR4A1; PDPK1;CAMK2A; IL2RG;CAMK2G
Downstream signaling of activated FGFR4	0.045	-2.235	2.684	NR4A1; PDPK1; CAMK2A;IL2RG; CAMK2G

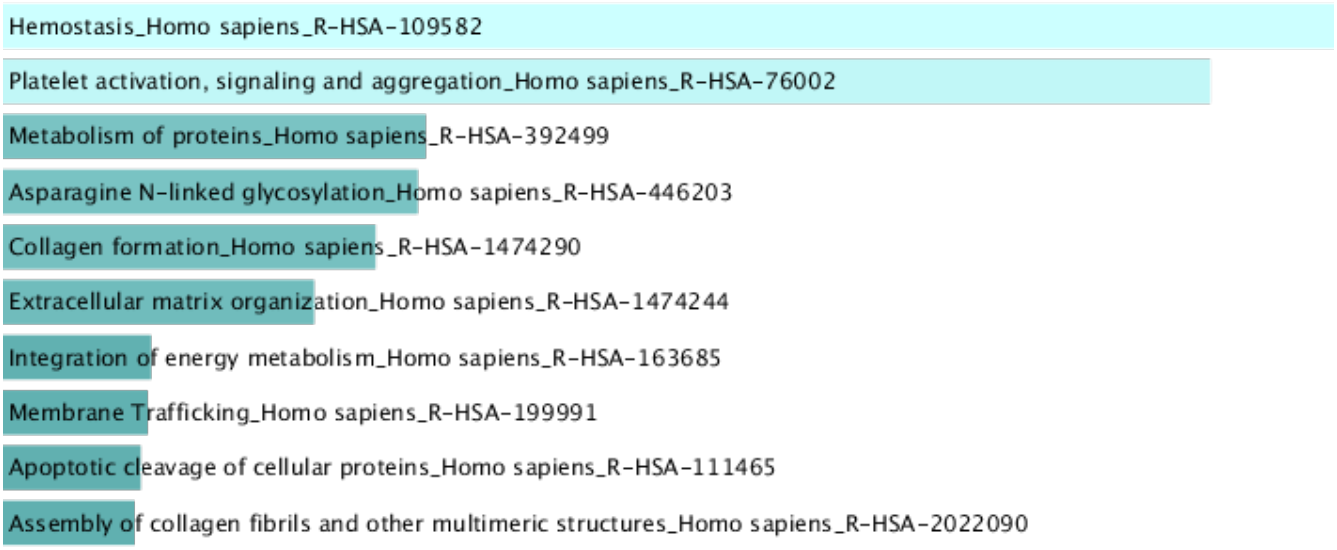

Figure 9: Enriched pathways associated with cancer gene fusion events in breast cancer as determined by Enrichr. Length of bar directly correlates to combined score as generated by Enrichr.

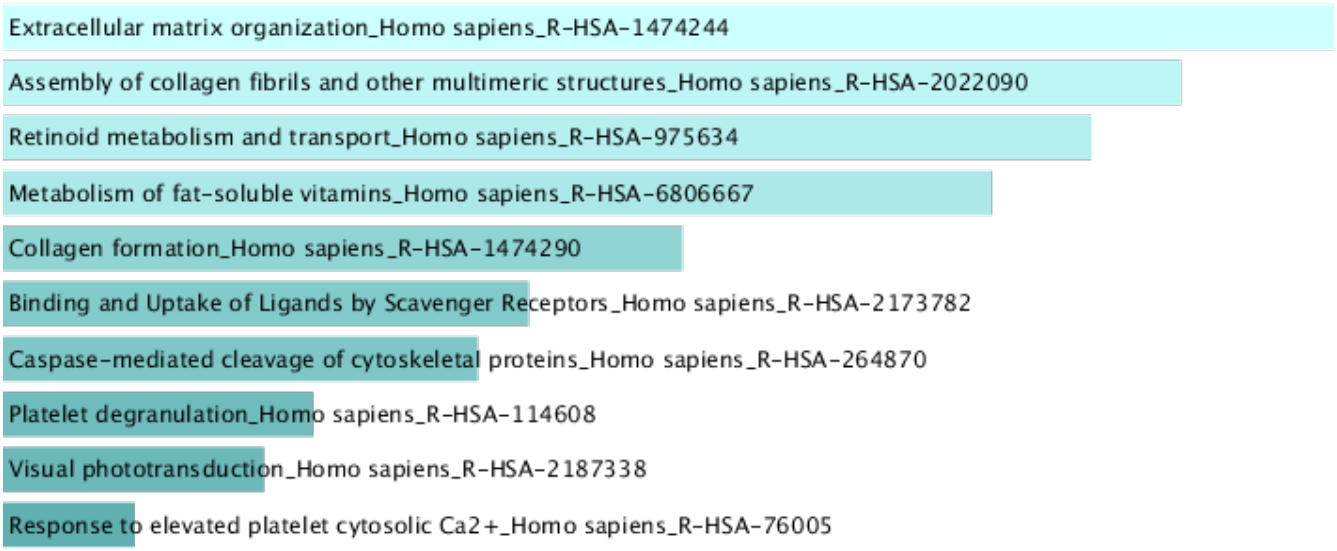

Figure 10: Enriched pathways associated with healthy gene fusion events in breast cancer as determined by Enrichr. Length of bar directly correlates to combined score as generated by Enrichr.

Table 7: Details of enriched pathways in gene fusions represented in breast cancer. P-value represents the p-value as determined by Fisher's exact test. Z-score is the z-score determined by the enrichr algorithm which is based on intuition and combined is the combined score between the p-value and the z-score.

Term	P-value	Z-score	Combined	Genes
Hemostasis	0.0000	-2.1481	13.8074	DGKG;LGALS3BP;APP;ITGAM;SPARC;SERPINE2;PIK3CD;KIF12;F11R; ITGAL;CLU;KIF5C;AKT2;PHACTR2;KIF21A;APLP2;SERPINF2;KIF25; ACTN4;MIF;GAB2;TUBA4A;SLC7A5;PRKAR1B;STIM1;KIFC1;KIF16B; SOS1;FN1;ALDOA;RAF1;DOCK2;MGLL;CD44;HDAC2;ITPR1;ITPR2; GATA3;PIK3R1;LRRC16A;PRKCZ;LRP8;GNAI2;RAP1B;RAP1A;GNG4; GNA12;ABL1;FLNA;STX4;FYN;CD74;CREBBP;HSPA5;NOS2;PCDH7;FN1; ATP2B2;PTK2;P2RX7;GNB2;GNAS;ESAM;GNB5;BCAR1
Platelet activation, signaling and aggregation	0.0000	-2.1525	12.7987	DGKG;LGALS3BP;APP;SPARC;ITPR1;ITPR2;PIK3CD;PIK3R1;PRKCZ; CLU;GNAI2;RAP1B;RAP1A;GNG4;AKT2;GNA12;FLNA;PHACTR2;STX4; FYN;HSPA5;APLP2;PCDH7;SERPINF2;FN1;ACTN4;GAB2;PTK2;TUBA4A; GNB2;GNB5;PFN1;ALDOA;RAF1;SOS1;MGLL;BCAR1
Metabolism of proteins	0.0002	-2.1278	6.4096	APP;ARF1;B4GALT1;PIGN;HSCB;ACTB;RPS15;TFG;RPL18A;KIF5C; TUBB3;PMPCA;RPL38;CTSD;SPTAN1;NUP214;EIF5B;ALG8;SEC13; NUP210;IGFBP5;MMP2;H2AFX;ALG12;ANK1;TUBA4A;DDOST;ARFGAP1; SUMF1;EEF1G;DPM1;CTAGE5;EEF1A2;RPL37A;EXOC3;NUP54;PABPC1; SEC22B;PFDN5;ARF5;SLC25A6;LTF;RAB1A;COPA;CPB1;TOMM40;GNAI2; GANAB;GYLTL1B;MTA1;GMDS;OS9;BGLAP;GNG4;LMAN2;NEU1;LMNA; RPL13;CD59;CCT7;RAE1;ST3GAL1;ST3GAL3;SPTBN2;GALNT7;FN3K; DYNC1H1;ARFGEF2;NOP58;XBP1;GSN;MUC16;HSPA5;EDEM1;FUCA1; PCGF2;TRAPPC10;STAT3;EIF2AK3;TIMM44;TRAPPC9;EEF2;GNAO1; MGAT4C;STAG2;EIF3L;GLB1;GNB2;GOLGB1;EIF3G;MSRB2;EIF3H;GNB5; TGFBI;EIF3D;EIF3A
Asparagine N-linked glycosylation	0.0002	-2.1059	6.3435	RAB1A;ARF1;COPA;B4GALT1;GANAB;GMDS;TFG;OS9;LMAN2;NEU1; CD59;ST3GAL1;SPTAN1;ST3GAL3;SPTBN2;DYNC1H1;ALG8;SEC13; EDEM1;FUCA1;TRAPPC10;ALG12;TRAPPC9;ANK1;ARFGAP1;DDOST; DPM1;MGAT4C;CTAGE5;GLB1;GOLGB1;SEC22B;ARF5
Collagen formation	0.0002	-1.9918	5.9997	LAMB3;COL22A1;PLOD1;MMP20;COL1A1;COL2A1;COL5A1;COL4A2; COL4A4;COL6A1;COL9A1;COL20A1;COL6A3;P4HB;CTSB;PLEC
Extracellular matrix organization	0.0004	-2.0721	5.4984	DDR1;FBN3;SPARC;PTPRS;ITGAM;ELN;PLOD1;F11R;ITGAL;MMP20; CAPN7;CDH1;NCAM1;CTSD;CTSB;LAMB3;COL22A1;MMP2;FN1;BGN; COL1A1;COL2A1;COL4A2;COL5A1;COL4A4;COL6A1;COL9A1;COL20A1; COL6A3;P4HB;AGRN;MATN4;CD44;PLEC
Integration of energy metabolism	0.0010	-1.9957	4.1710	PFKFB1;KCNJ11;ABCC8;ADIPOQ;ITPR1;CACNA1A;ITPR2;ACACA; GNAI2;CACNB3;RAP1A;PRKAR1B;GNG4;GNB2;GNAS;GNB5;SLC25A6
Membrane Trafficking	0.0014	-2.0586	4.1442	RAB1A;APP;ARF1;NBAS;COPA;TFRC;STX16;CLTB;KIF12;PIK3C2A;TFG; KIF5C;AKT2;LMAN2;SFN;STX4;KIF21A;CD59;SPTAN1;AP1M2;CYTH1; SPTBN2;DYNC1H1;SEC13;TRAPPC10;M6PR;KIF25;DTNBP1;AP3B1;SNF8; TRAPPC9;PUM1;ANK1;ARFGAP1;GJB2;CTAGE5;KIF16B;GOLGB1;EXOC3; TRIP11;SEC22B;ARF5;PICALM
Apoptotic cleavage of cellular proteins	0.0011	-1.9554	4.0866	OCLN;GSN;CDH1;LMNA;ACIN1;SPTAN1;PTK2;PLEC;LMNB1
Assembly of collagen fibrils and other multimeric structures	0.0012	-1.9317	4.0371	COL1A1;COL2A1;LAMB3;COL5A1;COL4A2;COL4A4;COL6A1;COL6A3; CTSB;PLEC;MMP20

Table 8: Details of enriched pathways in gene fusions represented in healthy breast tissue. P-value represents the p-value as determined by Fisher's exact test. Z-score is the z-score determined by the enrichr algorithm which is based on intuition and combined is the combined score between the p-value

and the z-score.

Term	P-value	Z-score	Combined	Genes
Extracellular matrix organization	5.47E-05	-2.072	10.351	COL28A1;FGB;COL18A1; LAMA3;LAMC2;LTBP3; HSPG2;FBLN5;MMP14; COL1A2;COL8A2;AGRN; CTSD;CTSB;PLEC
Assembly of collagen fibrils and other multimeric structures	3.97E-05	-1.990	10.202	COL18A1;COL1A2;LAMA3; COL8A2;LAMC2;CTSB;PLE C
Retinoid metabolism and transport	9.12E-06	-1.973	10.116	RBP4;LRP1;APOA1;LPL; BCO2;AGRN;HSPG2
Metabolism of fat-soluble vitamins	2.84E-05	-1.954	10.019	RBP4;LRP1;APOA1;LPL; BCO2;AGRN;HSPG2
Collagen formation	9.05E-05	-1.946	9.720	COL28A1;COL18A1;COL1A2 ;LAMA3;COL8A2;LAMC2;C TSB;PLEC
Binding and Uptake of Ligands by Scavenger Receptors	3.65E-05	-1.867	9.571	COL1A2;AMBP;LRP1;ALB; STAB1;SAA1;APOA1;FTL
Caspase-mediated cleavage of cytoskeletal proteins	8.65E-05	-1.906	9.521	GSN;MAPT;VIM;PLEC
Platelet degranulation	6.43E-05	-1.874	9.362	FGB;AHSG;APLP2;APOH; ALB;PSAP;IGF2;APOA1; CLU
Visual phototransduction	3.62E-05	-1.817	9.315	RBP4;GUCY2D;LRP1; RGS9BP;LPL;APOA1;BCO2; AGRN;HSPG2
Response to elevated platelet cytosolic Ca2+	8.99E-05	-1.840	9.189	FGB;AHSG;APLP2;APOH; ALB;PSAP;IGF2;APOA1; CLU

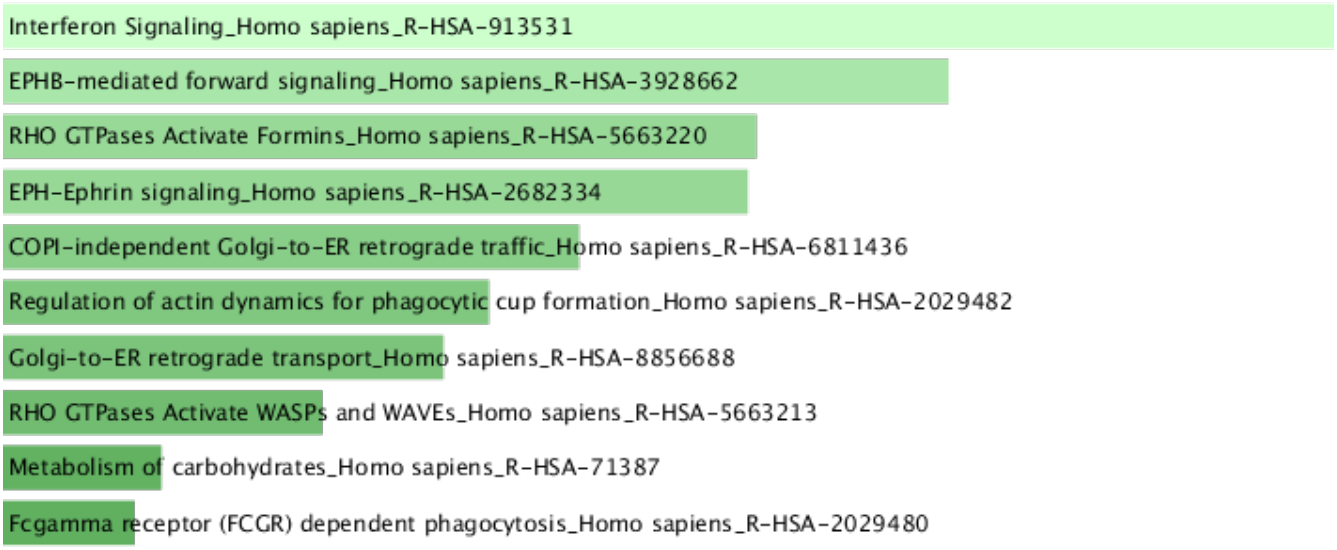

Figure 11: Enriched pathways associated with cancer gene fusion events in colorectal cancer as determined by Enrichr. Length of bar directly correlates to combined score as generated by Enrichr.

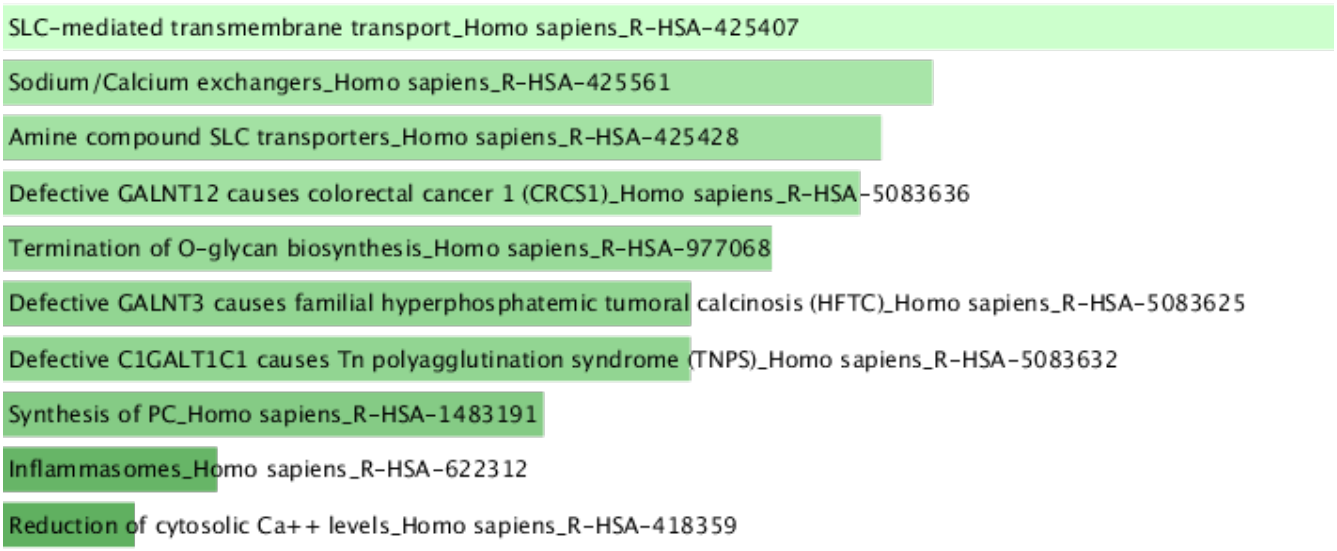

Figure 12: Enriched pathways associated with healthy gene fusion events in colorectal cancer as determined by Enrichr. Length of bar directly correlates to combined score as generated by Enrichr.

Table 9: Details of enriched pathways in gene fusions represented in colorectal cancer. P-value represents the p-value as determined by Fisher's exact test. Z-score is the z-score determined by the enrichr algorithm which is based on intuition and combined is the combined score between the p-value and the z-score.

Term	P-value	Z-score	Combined Score	Genes
Interferon Signaling	0.008	-2.078	2.065	GBP7;IFI27;GBP4;NUP160
EPHB-mediated forward signaling	0.013	-2.011	1.999	ACTB;ACTG1
RHO GTPases Activate Formins	0.011	-1.978	1.965	NUP160;ACTB;ACTG1
EPH-Ephrin signaling	0.007	-1.976	1.964	VAV3;ACTB;ACTG1
COPI-independent Golgi-to-ER retrograde traffic	0.007	-1.947	1.935	DCTN2;RAB6A
Regulation of actin dynamics for phagocytic cup formation	0.007	-1.931	1.919	VAV3;ACTB;ACTG1
Golgi-to-ER retrograde transport_Homo sapiens	0.010	-1.923	1.911	DCTN2;KIF13B;RAB6A
RHO GTPases Activate WASPs and WAVEs	0.010	-1.902	1.891	ACTB;ACTG1
Metabolism of carbohydrates	0.026	-2.121	1.863	PYGB;AKR1A1;HAS2;NUP160
Fcgamma receptor (FCGR) dependent phagocytosis	0.013	-1.870	1.858	VAV3;ACTB;ACTG1

Table 10: Details of enriched pathways in gene fusions represented in healthy colon tissue. P-value represents the p-value as determined by Fisher's exact test. Z-score is the z-score determined by the enrichr algorithm which is based on intuition and combined is the combined score between the p-value and the z-score.

Term	P-value	Z-score	Combined	Genes
SLC-mediated transmembrane transport	0.012	-1.983	5.452	SLC44A1;SRI
Sodium/Calcium exchangers	0.008	-1.903	5.232	SRI
Amine compound SLC transporters	0.021	-2.012	5.204	SLC44A1
Defective GALNT12 causes colorectal cancer 1 (CRCS1)	0.013	-1.888	5.192	MUC13
Termination of O-glycan biosynthesis_Homo sapiens	0.018	-1.965	5.144	MUC13
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	0.013	-1.855	5.099	MUC13
Defective C1GALT1C1 causes Tn polyagglutination syndrome (TNPS)	0.014	-1.855	5.099	MUC13
Synthesis of PC	0.014	-1.825	5.018	SLC44A1
Inflammasomes	0.012	-1.760	4.839	PANX1
Reduction of cytosolic Ca++ levels	0.009	-1.744	4.794	SRI

4.5 GENE EXPRESSION ANALYSIS

From our breast cancer classification, it was found that accuracy, precision, recall and F-score decreased when fusion events from TNBC and ER+ types were combined. As a result, we hypothesized the regulation patterns surrounding either of the gene fusion partners' pathways would be affected. We also hypothesized that this may be a good feature for our gene fusion classification model.

To test this hypothesis, we found the Reactome pathway neighbors within a distance of two away from either fused gene using PathwayCommons. This takes into account genes directly upstream and downstream of the gene on the same pathway. However, this also takes into account genes that are direct neighbors of genes that are directly adjacent to the gene fusion partners. This method allows us to explore the network surrounding the pathway more efficiently and more closely look at how the gene fusion affects gene expression for genes that are not in the same pathway.

After the gene neighbors are determined for both cancer and healthy gene fusions, their expression values are detected using the Tuxedo Suite. Tophat (Trapnell, 2009) was used to align samples to the genome and Cufflinks (Trapnell, 2010; Trapnell, 2013) was used to quantify read counts and normalize based on read length. ΔFPKM was detected using the following formula for neighbors of healthy and cancer gene fusion events in each cancer studied:

$$\Delta\text{FPKM} = \text{FPKM}_{normal} - \text{FPKM}_{cancer}$$

From here the FPKM is converted into a z-score in order to further normalize the samples. Standard deviation and mean values are calculated for the healthy samples and then applied to the following formula for both cancer and healthy neighbors. For cancer patients, the z-score was calculated using the healthy standard deviation and mean, in order to compare the two groups to see where the cancer was on the normal curve:

$$z = \frac{x - \mu_{normal}}{\sigma_{normal}}$$

Where z is the z-score, x is the delta FPKM of a particular gene, μ is the mean of the non-

malignant gene fusion neighbors and σ us the standard deviation of the non-malignant fusion neighbors. To find the most differentially expressed of these samples, the genes were filtered so that only genes with a z-score of > 3 or < -3 were present and differential expression was measured.

Differentially expressed neighbors of gene fusion partners were mapped to histograms to study the density. From here, we can determine that the density of healthy and cancer gene fusion neighbors is the same for each type of cancer, as expected (Figures 13-15).

We looked at all gene fusions that we associated with cancers and tried to find their overlap in their pathway neighbors, determined as mentioned above (Figure 16). Amongst all of the cancers there were two genes that were common amongst all of the cancers, *B2M* and *PABPC1*. *B2M* is a gene that codes for a serum protein found In the MHC class I heavy chain. It is also known to be associated with colon cancer (Kheirelseid, 2010). *PABPC1* is a gene which encodes polyadenylate binding protein 1 which binds the poly-A tail of the mRNA (Grosset, 2000). An interesting finding that might need to be studied further is that the expression values for our representative hematological cancer (ALL) has a different gene expression pattern for both of these genes than our two representative solid tumor cancers (BC, and CRC). *B2M* is upregulated in ALL, while downregulated in both solid tumors, and *PABPC1* is downregulated in ALL, while upregulated in both solid tumors. While we only studied three cancers, it might be interesting to look at the differences between other hematological cancers and solid tumors.

Next, we were interested in finding the unique neighbor genes of gene fusion events in each respective cancer. We took the set difference between the unique neighbor genes in healthy and the unique neighbor genes in cancer. It is hypothesized that the unique gene fusion neighbors will be associated with cancer pathways or have functions related to cancer. Since we did not specify whether these genes will be up or down regulated (we took both z-scores greater than 3 and less than -3) and the genes can come from anywhere, we believe that if these neighbor genes are associated with cancer pathways, then these fusion neighbors might be good therapeutic targets.

Acute Lymphoblastic Leukemia had only one unique differentially expressed neighbor for its cancer (Figure 17) and healthy (Figure 18) gene fusion events. The unique fusion neighbor present in Acute Lymphoblastic Leukemia is *NPM1*. *NPM1* is involved in the regulation of the *ARF*/*p53* pathway. It is also involved in nucleic acid binding and protein homodimerization, according to its Gene Ontology terms. It is known as a frequently mutated gene in hematological cancers (Balusu, 2011). This gene is upregulated in our cancer dataset. The unique fusion neighbor present in healthy human cord blood is *CALR*. This gene is a member of the *JAK/STAT* pathway, which while known for its diabetes signaling capabilities, its members have also showed up in cancer pathways. *JAK* is a tyrosine kinase, which members are frequently associated with fusion events and the *JAK/STAT* pathway is commonly affected in cancer. Its role in normal cells is to serve as an "eat-me" sign for phagocytosis (Lu, 2014). This gene is downregulated in our cancer dataset.

For our unique gene fusion pathway neighbors in breast cancer (Figure 19) three genes are interesting, *GAPDH*, *MGST1* and *ACTA2*. *GAPDH*, despite its use as a housekeeping gene, is actually well known to have increased expression in certain types of solid cancers (Tokunaga, 1987). SNPs for *MGST* are found in colorectal cancer. SNPs in this gene could contribute to the risk in young people (< 50 years old) (Iida, 2001). *ACTA2* encodes Alpha smooth muscle actin 2, and studies have shown its connection with early brain metastasis of lung adenocarcinoma. For our unique gene fusion pathway neighbors in healthy breast tissue (Figure 20), some of the interesting genes that are related with cancer are *GPX3* and *RBP4*. *GPX3* is a in the glutathione peroxidase family and shown to be downregulated in cancer (Zhang, 2010). It is also required for synthesis and storage of intracellular triglycerides. *RBP4*, or Retinol Binding Protein 4, is also involved in the *JAK/STAT* pathway. *RBP4* isoforms have also been shown to be decreased in liver cancer (Frey, 2008).

Finally, for our unique gene fusion pathway neighbors in colorectal cancer (Figure 21), some of the interesting genes that are related with cancer include *CCT3*, *HMGA1*. *CCT3* is a chaperone protein and assists with the folding of proteins upon ATP hydrolysis. It is also associated with hepatocellular

carcinoma. *HMGA1* is a gene that is associated with tumor progression in TNBC breast cancer cells (Shah, 2013). Three genes were found as unique for our gene fusion pathway neighbors in healthy colon tissue (Figure 22), *GCNT3*, *SLC44A4* and *SRI*. Both *SLC44A4* and *SRI* are signaling genes, with *SLC44A4* being a sodium-dependent transport protein and *SRI* being a calcium binding protein. *GCNT3* is actually a marker for colon cancer and lower expression indicates higher probability of disease-free survival (Gonzalez-Vallinas, 2014).

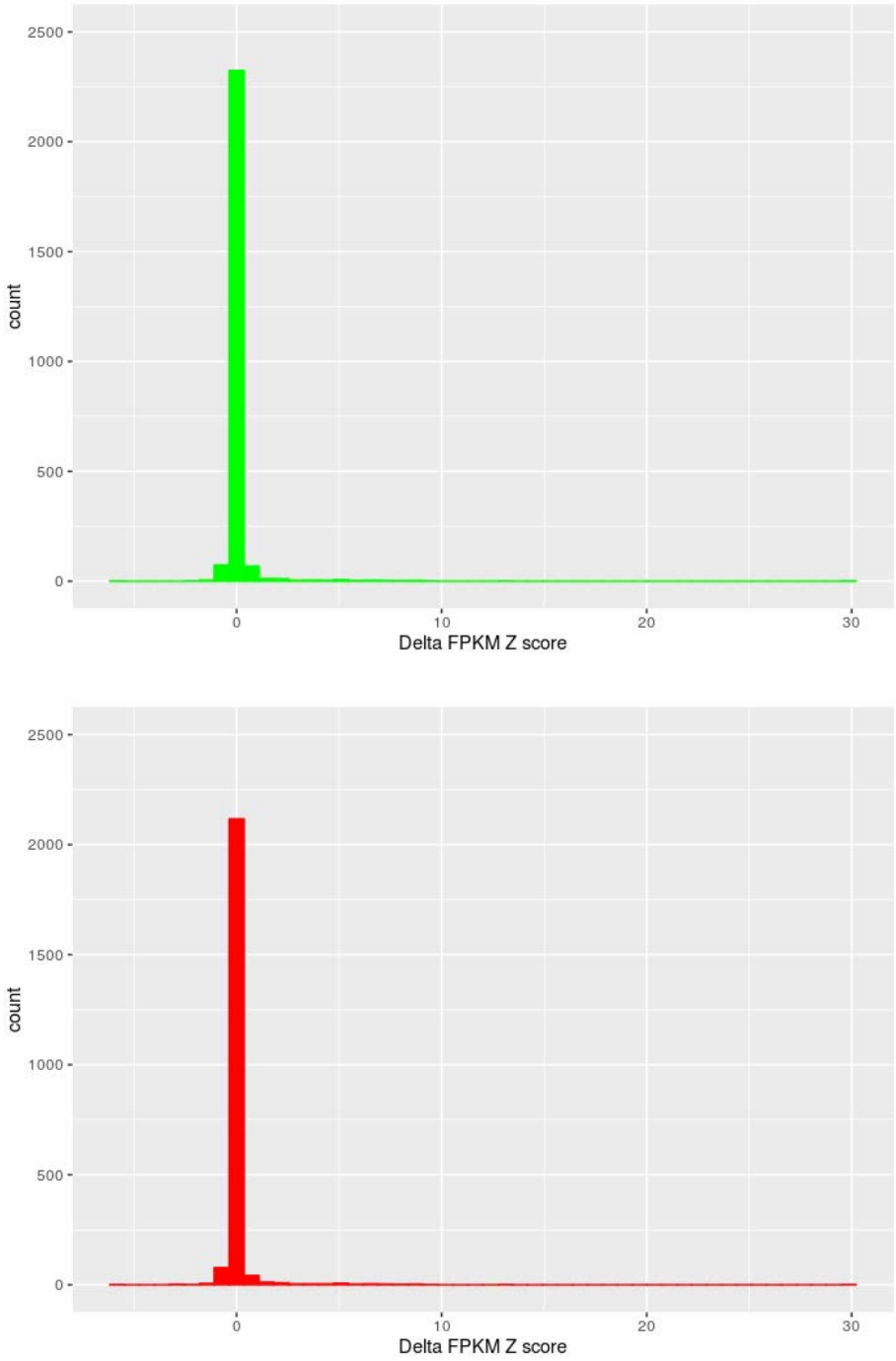

Figure 13: Delta FPKM of acute lymphoblastic leukemia fusion neighbor densities healthy (top) and cancer (bottom).

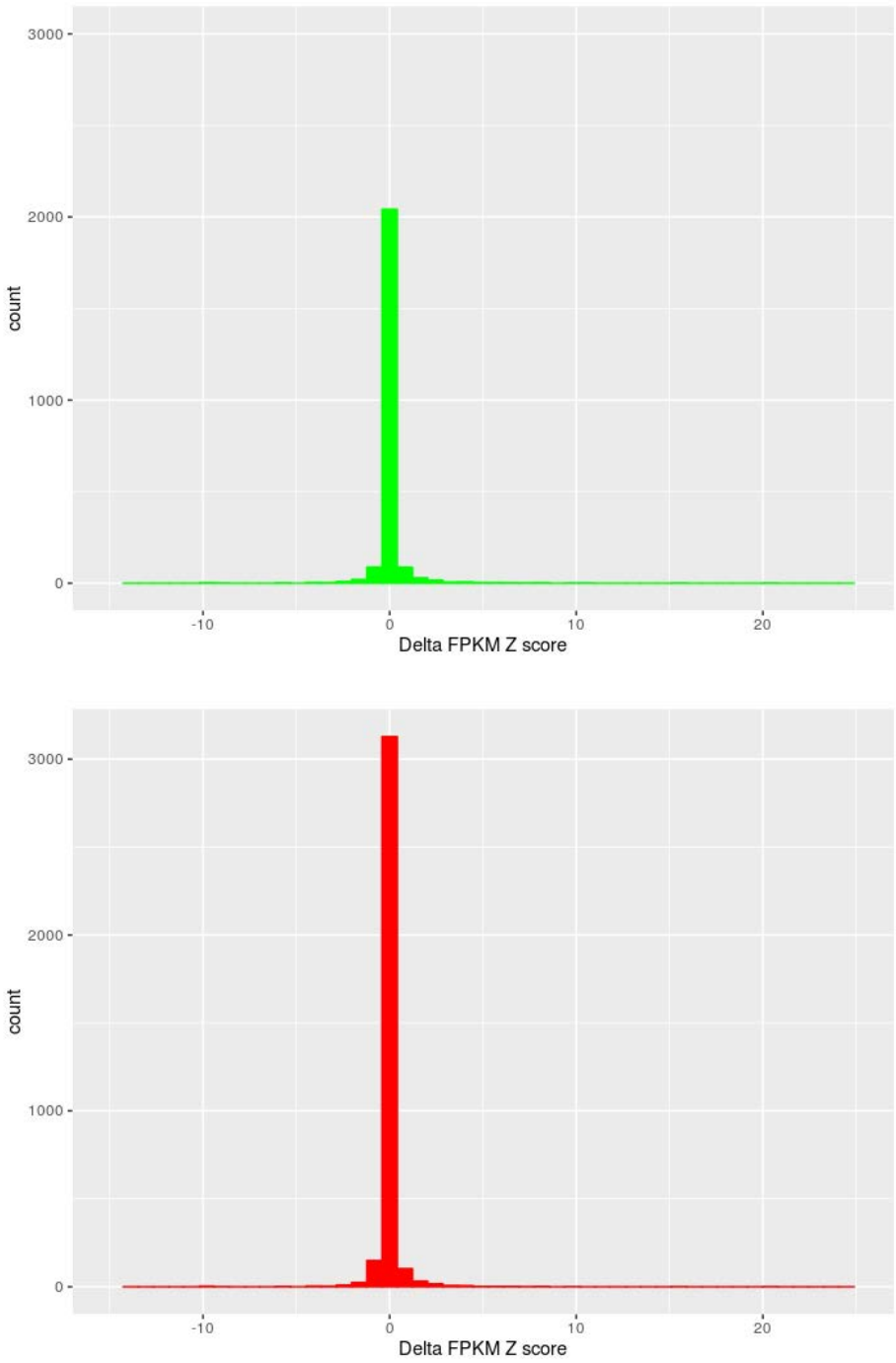

Figure 14: Delta FPKM of breast cancer fusion neighbor densities healthy (top) and cancer (bottom).

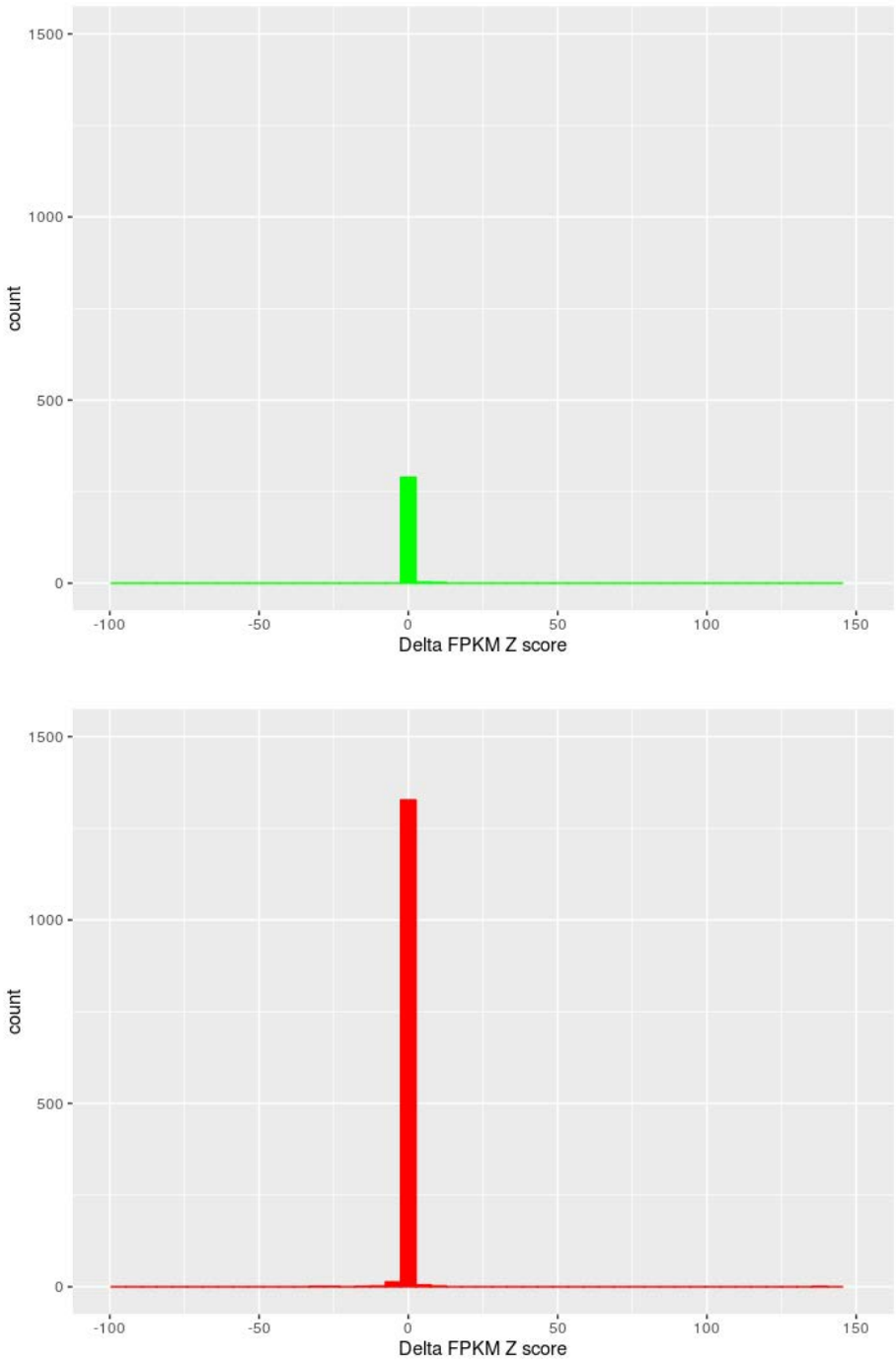

Figure 15: Delta FPKM of colorectal cancer fusion neighbor densities healthy (top) and cancer (bottom).

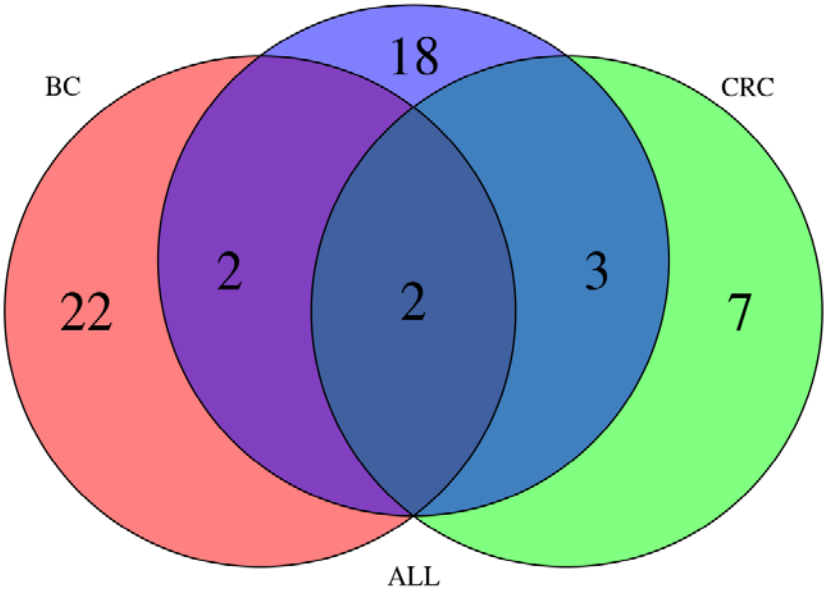

Figure 16: Venn diagram of all of the overlapping neighbors between cancer fusion neighbors in ALL (blue), breast cancer (pink) and colorectal cancer (green)

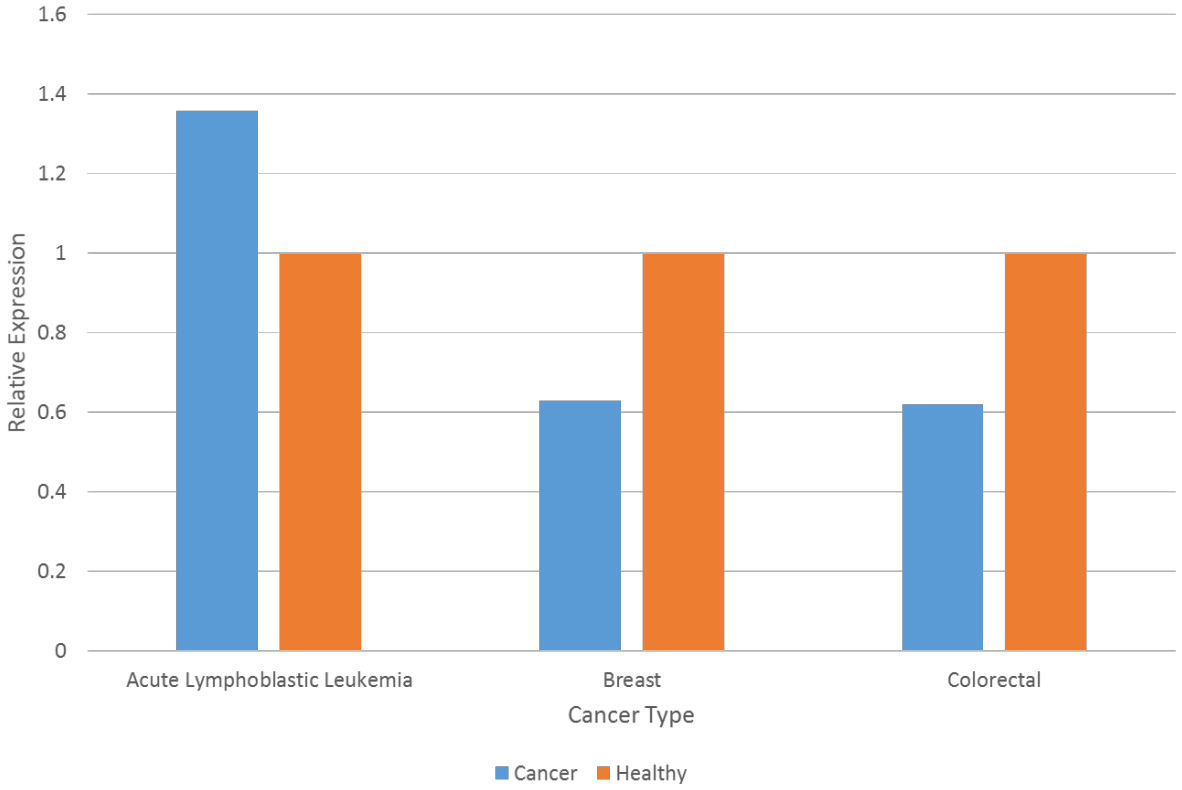

Figure 17: Relative expression of B2M, which is a common gene fusion neighbor between colon cancer, acute lymphoblastic leukemia and breast cancer

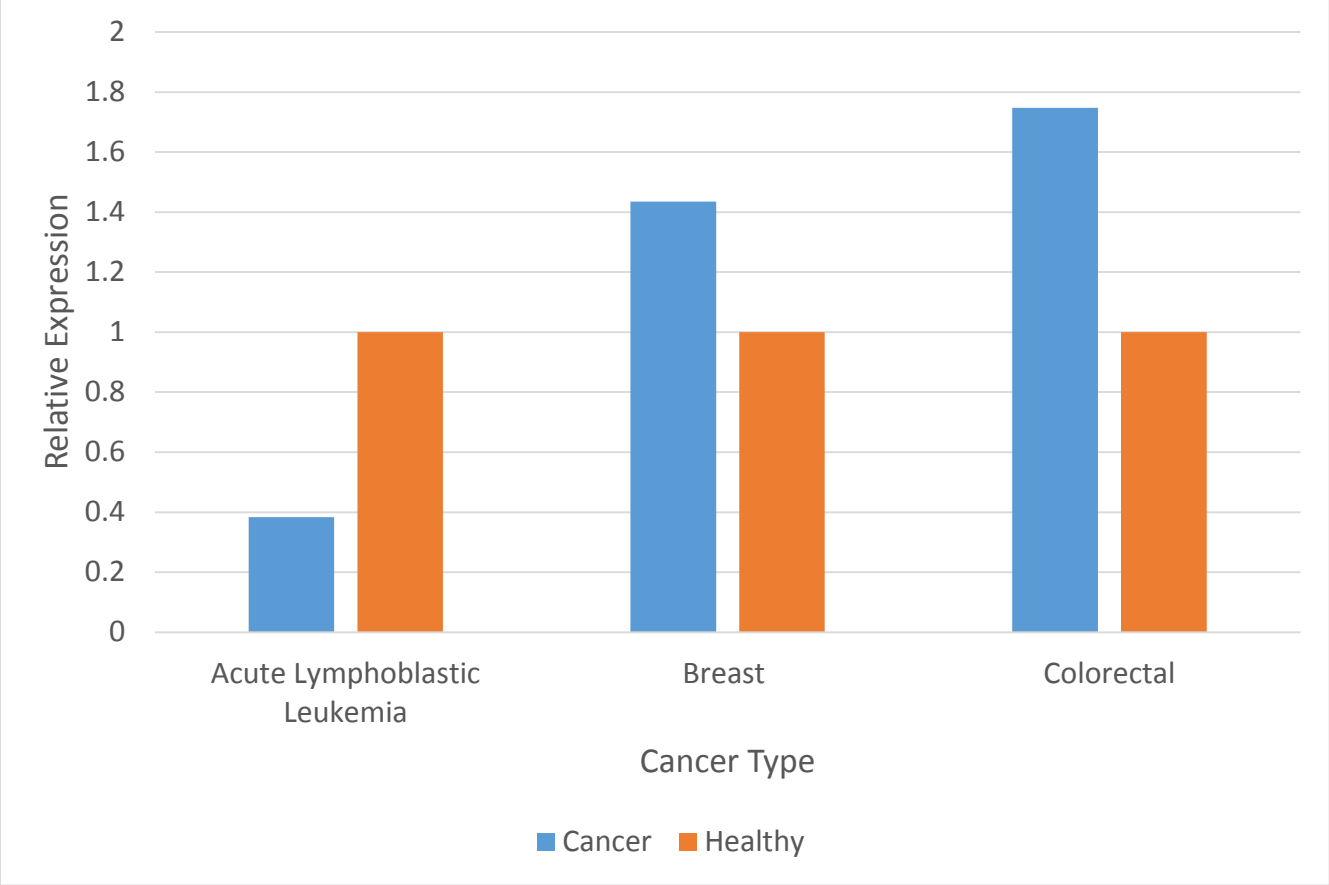

Figure 18: Relative expression of *PABPC1*, which is a common gene fusion neighbor between colon cancer, acute lymphoblastic leukemia and breast cancer

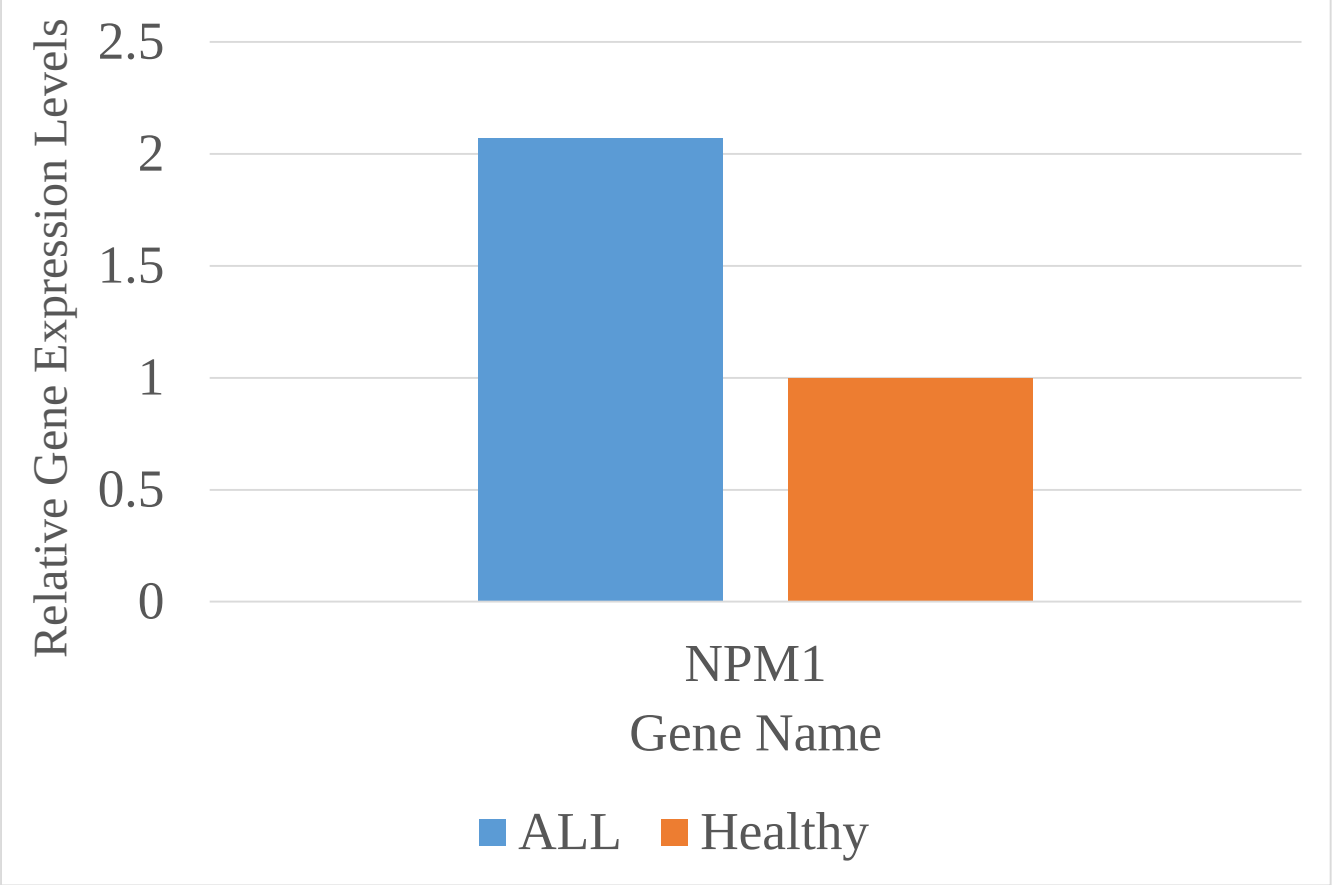

Figure 19: Acute lymphoblastic leukemia unique in cancer fusion neighbors relative gene expression values for both cancer (blue) and healthy (orange).

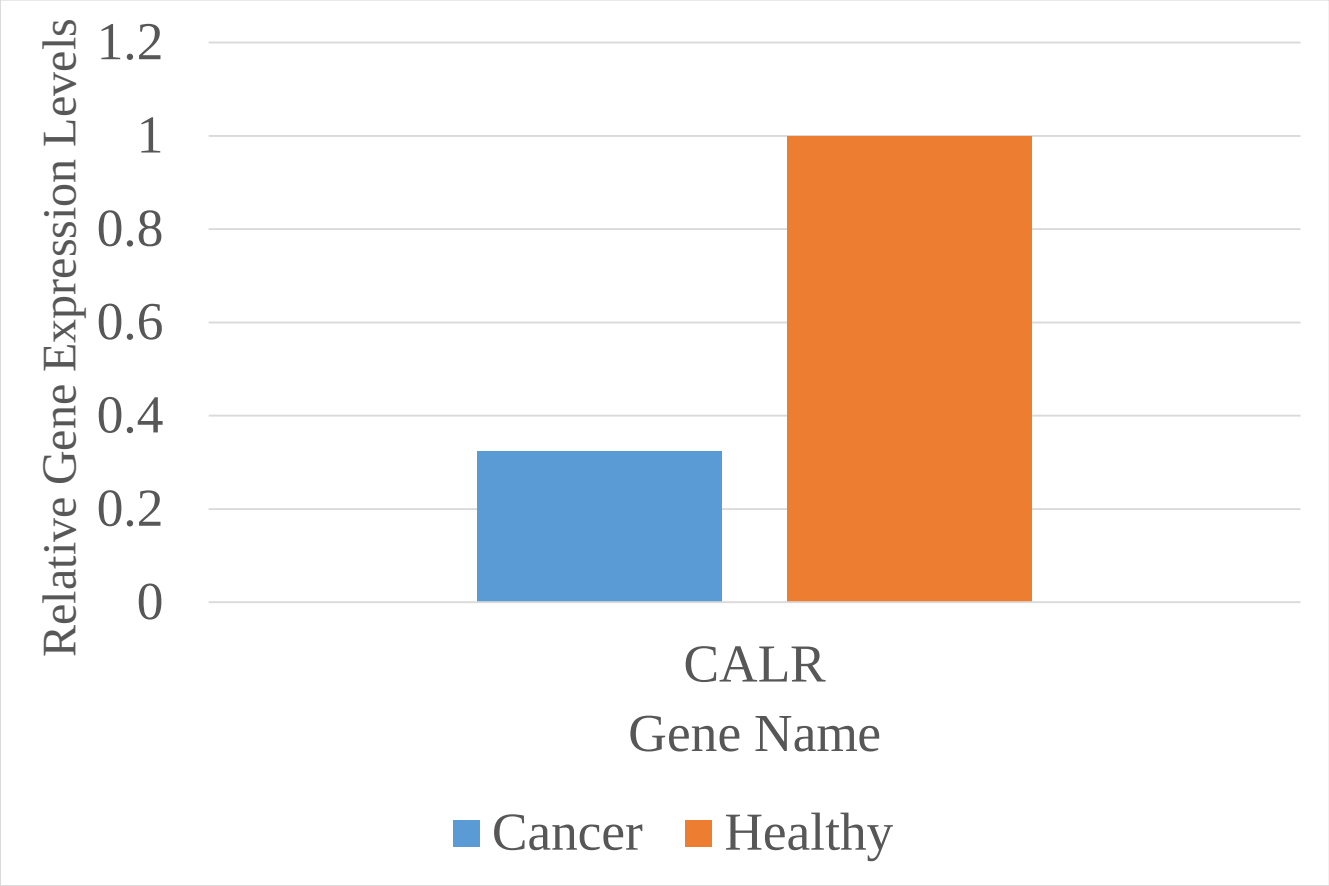

Figure 20: Acute lymphoblastic leukemia unique in healthy fusion neighbors relative gene expression values for both cancer (blue) and healthy (orange).

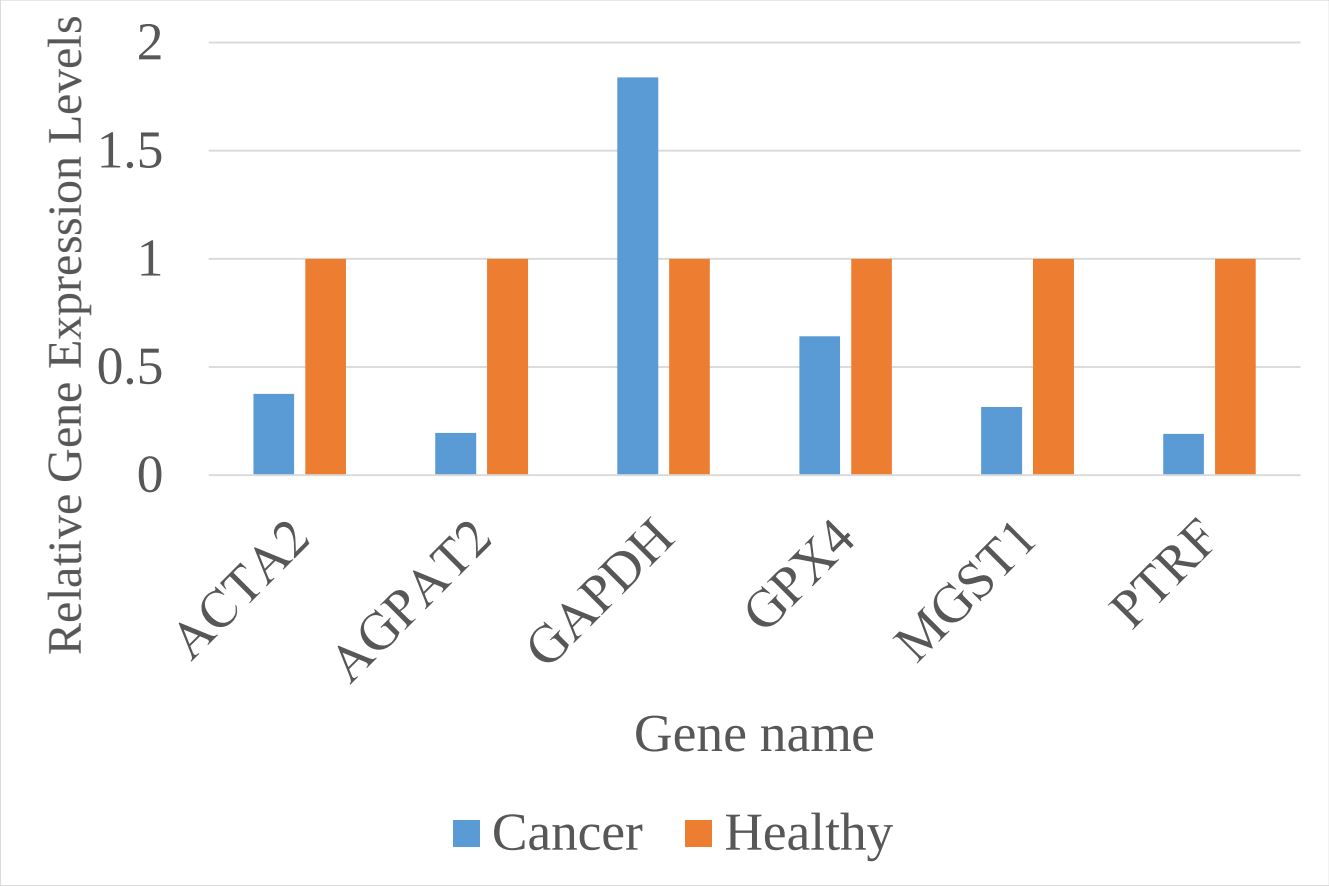

Figure 21: Breast cancer unique in cancer fusion neighbors relative gene expression values for both cancer (blue) and healthy (orange).

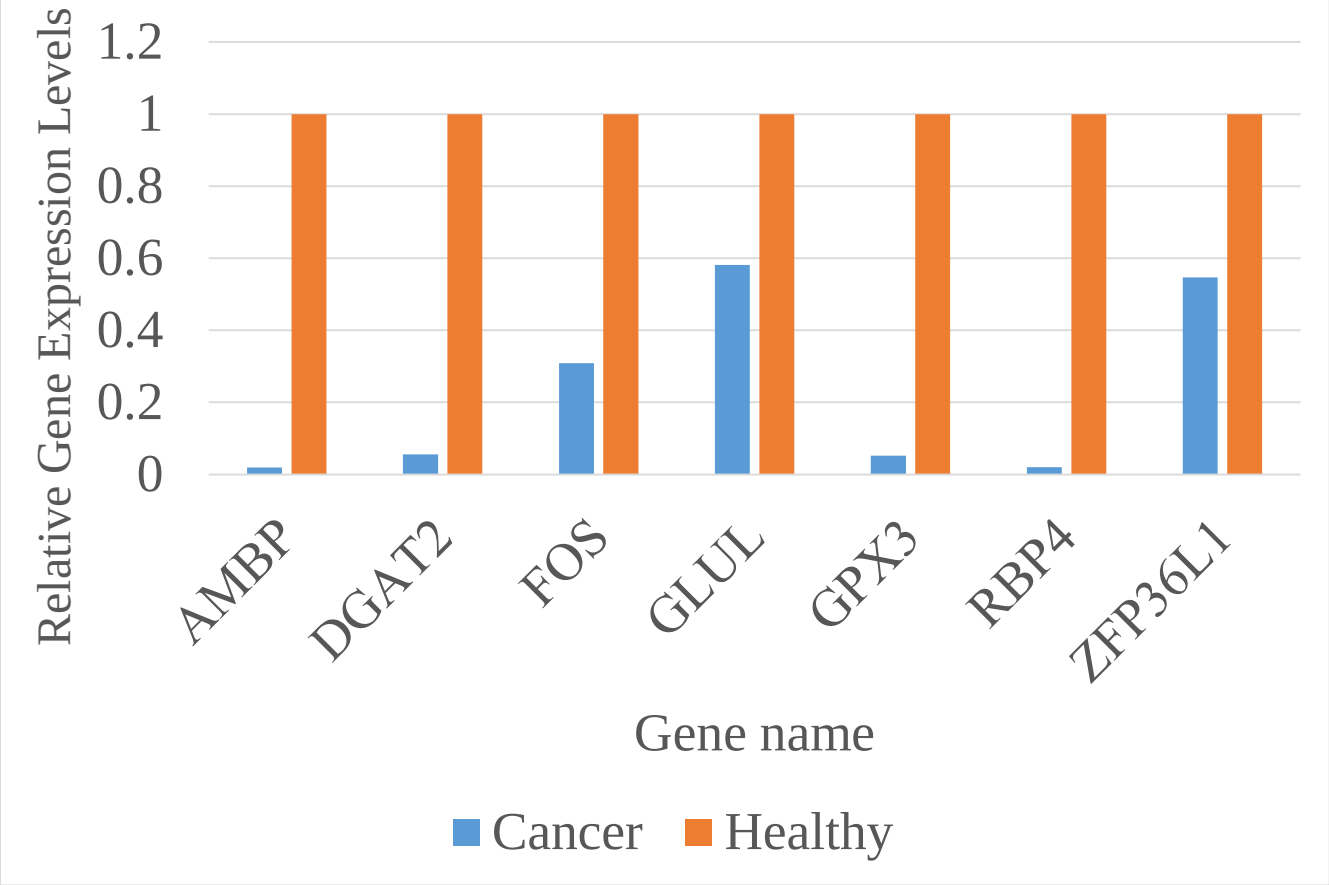

Figure 22: Breast tissue unique in healthy fusion neighbors relative gene expression values for both cancer (blue) and healthy (orange).

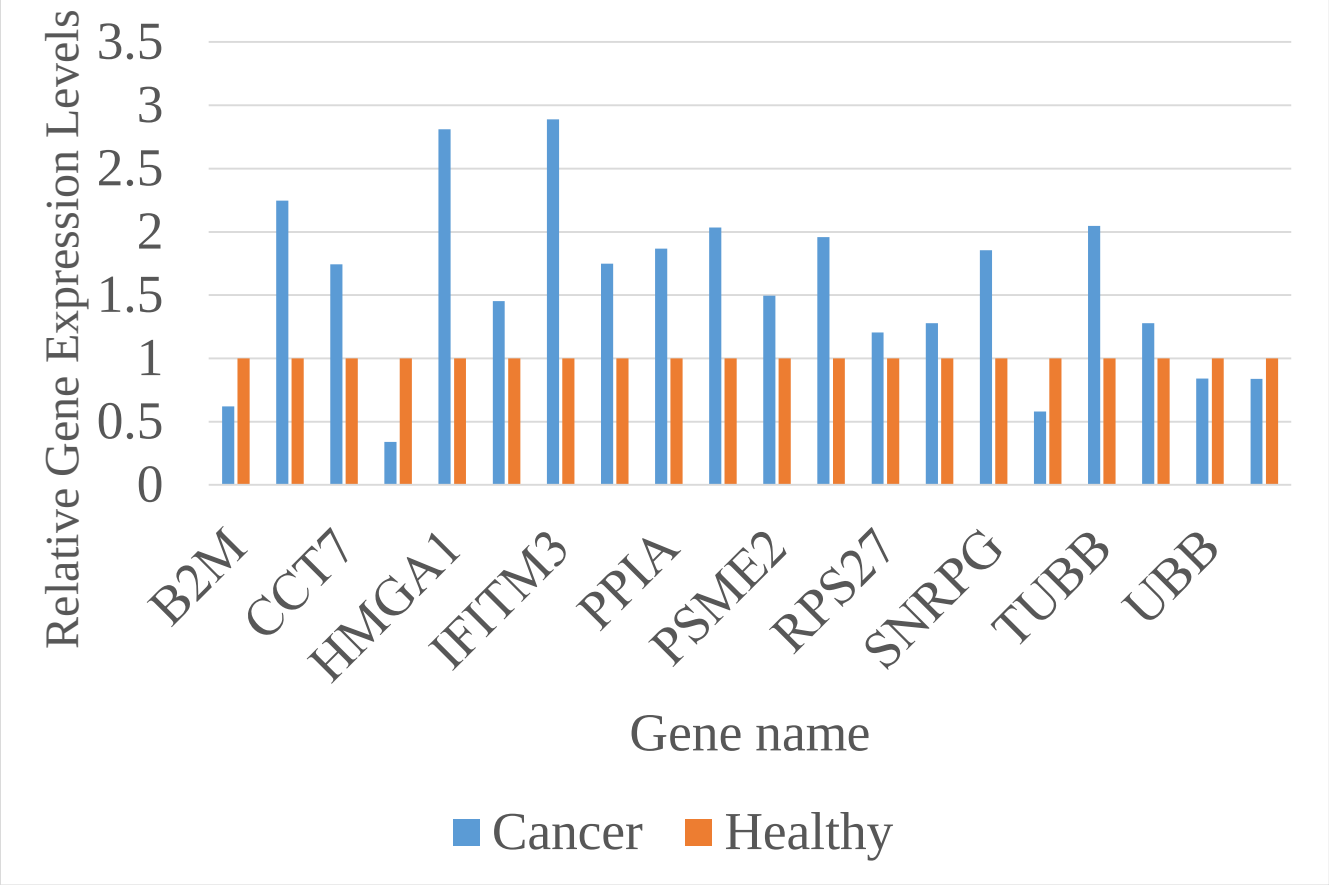

Figure 23: Colorectal cancer unique in cancer fusion neighbors relative gene expression values for both cancer (blue) and healthy (orange).

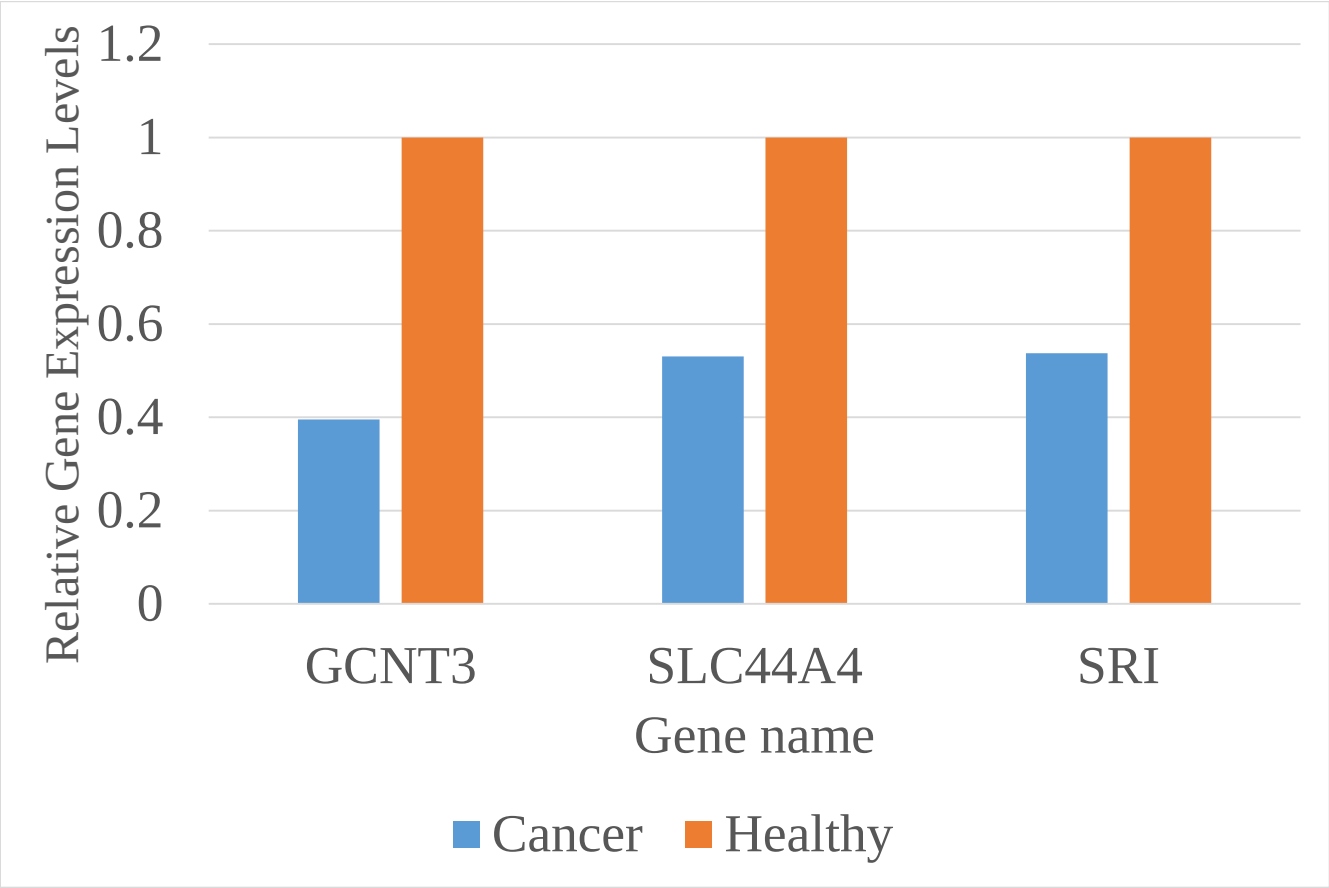

Figure 24: Colorectal unique in cancer fusion neighbors relative gene expression values for both cancer (blue) and healthy (orange).

5. DISCUSSION

Since the start of our current age of genomics and next-generation sequencing, new mutations are able to be discovered and confirmed more quickly. SNVs and CNVs have been studied in depth through large scale GWAS studies and other means, but gene fusion detection and analysis is still inconclusive. The deluge of gene fusion detection software programs leaves us with more questions than answers at this point, especially because there are over 30 different software programs and most have been shown through independent studies as having little overlap. The purpose of our study was two-fold: (1) Accurately classify gene fusion events belonging to cancer and healthy tissue or cells and (2) characterize the gene fusion pathway landscape and garner better insight into how gene fusions might drive cancer.

Through this study, we were able to accurately classify gene fusion events derived from FusionMap as belonging to the cancer or healthy samples, despite the fact that many overlapping gene fusion events between cancer and healthy samples were found. More analyses need to be done in order to determine if semi-supervised learning is the best model for gene fusion data, since semi-supervised models performed similar to supervised models after performing feature selection. Modifying our features to include gene expression information of pathway neighbors may improve our model. Our model does solve the problem of determining a tissue-specific model for classifying gene fusion events, since we observe that there are no overlapping genes between any of the genes involved in gene fusions or their neighbors.

Through exploring the gene fusion pathway neighbors we were able to observe many gene cancer related pathways affected, as represented by the functions of each of the unique pathway neighbor genes of gene fusion events found as having z-scores of greater than 3 or less than -3. Each type of tissue, whether the neighbor was unique to cancer or healthy, contained at least one gene that was within a cancer pathway. This predicts the existence of the transcription ripple effect caused by gene fusions to the rest of their pathway networks. Further research and possible collaboration with a

wet lab would be needed to see under which circumstances this holds true.

Not all of the gene fusion neighbors found were prevalent in cancer pathways, so it is important to further look into the literature to find a connection between these genes and their fusions. This includes looking into whether the gene fusions have frame-shift mutations, intact domains or produce a functional gene product.

Our long-term goal is to represent our pathway neighbor analysis as a feature in our algorithm. It is predicted that this will improve our algorithm, as it seems like cancer pathways are overall affected in gene fusion events. In order to do this, we would also have to test a normal dataset in order to see if there is a significant difference between our findings and non-fused gene neighbors in cancer. We would also like to test our current model with an independent healthy sample from another lab, which would make sure our model is not biased based on lab.

In order to improve our pathway neighbor analysis, we would like to connect the pathway neighbors found back to the gene fusion events that they are associated with. Can we find a direct connection between these genes and their fused neighbors? Currently, both upstream and downstream effects are grouped together. It might be interesting to isolate upstream and downstream effects to compare if downstream effects have greater differential expression than upstream effects.

Kinases and transcription factors are heavily involved in gene fusion events. It is hypothesized, if our gene fusions detected from FusionMap are not false positives, that a significant portion of them will contain transcription factors or kinases. It would be useful to count these groups in each of the datasets used, and separate into cancer and healthy to observe differences.

Overall in this study, we developed a model able to predict fusion events associated with cancer and healthy phenotypes with about an 85% accuracy. We have also found that pathway neighbors of gene fusion events that are significantly differentially expressed have direct connections to cancer pathways. These neighbors could be used to further improve therapeutic targets for cancer, especially those cancers known to strongly be affected by gene fusions.

6. REFERENCES

1. Abate, F. et al. Bellerophontes: An RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. Bioinformatics 28, 2114–2121 (2012).

2. Abate, F. et al. Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. BMC Syst. Biol. 8, 97 (2014).

3. Almamun, M. et al. Integrated methylome and transcriptome analysis reveals novel regulatory elements in pediatric acute lymphoblastic leukemia. Epigenetics 10, 882–890 (2015).

4. Asmann, Y. W. et al. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res. 39, (2011).

5. Beccuti, M. et al. Chimera: A Bioconductor package for secondary analysis of fusion products. Bioinformatics 30, 3556–3557 (2014).

6. Ben-Hur, A., Ong, C. S., Sonnenburg, S., Scholkopf, B. & Rutsch, G. Support vector machines and kernels for computational biology. PLoS Comput. Biol. 4, (2008).

7. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

8. Breiman, L. Random forests. Mach. Learn. 45, 5–32 (2001).

9. Carrara, M. et al. State-of-the-art fusion-finder algorithms sensitivity and specificity. Biomed Res. Int. 2013, (2013).

10. Cerami, E. G. et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 39, 685–690 (2011).

11. Cui, H., Dhroso, A., Johnson, N. & Korkin, D. The variation game: Cracking complex genetic disorders with NGS and omics data. Methods 79, 18–31 (2015).

12. Davidson, N. M., Majewski, I. J. & Oshlack, A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. Genome Med. 7, 43 (2015).

13. Demichelis, F. et al. TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. Oncogene 26, 4596–4599 (2007).

14. Engreitz, J. M., Agarwala, V. & Mirny, L. A. Three-Dimensional Genome Architecture Influences Partner Selection for Chromosomal Translocations in Human Disease. PLoS One 7, 1–9 (2012).

15. Fernandez-Cuesta, L. et al. Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. Genome Biol. 16, 7 (2015).

16. Francis, R. W. et al. Fusionfinder: A software tool to identify expressed gene fusion candidates from RNA-seq data. PLoS One 7, (2012).

17. Ge, H. et al. FusionMap: Detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics 27, 1922–1928 (2011).

18. Hall, M. a. Correlation-based Feature Selection for Machine Learning. Methodology 21i195-i20, 1–5 (1999).

19. Iyer, M. K., Chinnaiyan, A. M. & Maher, C. A. ChimeraScan: A tool for identifying chimeric transcription in sequencing data. Bioinformatics 27, 2903–2904 (2011).

20. Janz, S., Potter, M. & Rabkin, C. S. Lymphoma- and leukemia-associated chromosomal translocations in healthy individuals. Genes Chromosom. Cancer 36, 211–223 (2003).

21. Jia, W. et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol. 14, R12 (2013).

22. Kanehisa, M. & Goto, S. Yeast Biochemical Pathways. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27–30 (2000).

23. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462 (2016).

24. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12, R72 (2011).

25. Kim, S. K. et al. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. Mol. Oncol. 8, 1653–1666 (2014).

26. Kumar-Sinha, C., Kalyana-Sundaram, S. & Chinnaiyan, A. M. Landscape of gene fusions in epithelial cancers: seq and ye shall find. Genome Med. 7, 129 (2015).

27. Liu, J. et al. Diagnostic Pathology : Open Access Fusion Genes and Their Detection through Next Generation Sequencing in Malignant Hematological Diseases and Solid Tumors. 1, 1–6 (2016).

28. McPherson, A. et al. Defuse: An algorithm for gene fusion discovery in tumor rna-seq data. PLoS Comput. Biol. 7, (2011).

29. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 15, 371–381 (2015).

30. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. Nat. Rev. Cancer 7, 233–45 (2007).

31. Mullighan, C. G. The molecular genetic makeup of acute lymphoblastic leukemia. Hematology Am. Soc. Hematol. Educ. Program 2012, 389–96 (2012).

32. Pakakasama, S. et al. Simple multiplex RT-PCR for identifying common fusion transcripts in childhood acute leukemia. Int. J. Lab. Hematol. 30, 286–291 (2008).

33. Seeger, K. et al. TEL-AML1 fusion transcript in relapsed childhood acute lymphoblastic leukemia. The Berlin-Frankfurt-Münster Study Group. Blood 91, 1716–1722 (1998).

34. Shugay, M., De Mendíbil, I. O., Vizmanos, J. L. & Novo, F. J. Oncofuse: A computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics 29, 2539–2546 (2013).

35. Shugay, M., Ortiz de Mendíbil, I., Vizmanos, J. L. & Novo, F. J. Genomic Hallmarks of Genes Involved in Chromosomal Translocations in Hematological Cancer. PLoS Comput. Biol. 8, (2012).

36. Soda, M. et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 448, 561–566 (2007).

37. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. Nat. Commun. 5, 4846 (2014).

38. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 31, 46–53 (2013).

39. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111 (2009).

40. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511 (2010).

41. Vapnik, V. & Vashist, A. A new learning paradigm: Learning using privileged information. Neural Networks 22, 544–557 (2009).

42. Varley, K. E. et al. Recurrent read-through fusion transcripts in breast cancer. Breast Cancer Res. Treat. 146, 287–297 (2014).

43. Wang, Q., Xia, J., Jia, P., Pao, W. & Zhao, Z. Application of next generation sequencing to human gene fusion detection: Computational tools, features and perspectives. Brief. Bioinform. 14, 506–519 (2013).

44. Yoshihara, K. et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene 1–10 (2014). doi:10.1038/onc.2014.406

45. Zelent, A., Greaves, M. & Enver, T. Role of the TEL-AML1 fusion gene in the molecular pathogenesis of childhood acute lymphoblastic leukaemia. Oncogene 23, 4275–4283 (2004).

46. Haraksingh, R. R. & Snyder, M. P. Impacts of variation in the human genome on gene regulation. J. Mol. Biol. 425, 3970–3977 (2013).

47. Song, J., Mercer, D., Hu, X., Liu, H. & Li, M. M. Common leukemia- and lymphoma-associated genetic aberrations in healthy individuals. J. Mol. Diagnostics 13, 213–219 (2011).

48. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128 (2013).

49. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. & Ebert, B. L. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. Proc Natl Acad Sci U S A 102, 15545–15550 (2005).

50. Djebali, S. et al. Evidence for transcript networks composed of chimeric rnas in human cells. PLoS One 7, (2012).

51. Parra, G. et al. Tandem chimerism as a means to increase protein complexity in the human genome Tandem chimerism as a means to increase protein complexity in the human genome. 37–44 (2006). doi:10.1101/gr.4145906

52. Akiva, P. et al. Transcription-mediated gene fusion in the human genome Transcription-mediated gene fusion in the human genome. 30–36 (2006). doi:10.1101/gr.4137606

53. Casado-Vela, J., Lacal, J. C. & Elortza, F. Protein chimerism: Novel source of protein diversity in humans adds complexity to bottom-up proteomics. Proteomics 13, 5–11 (2013).

54. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res. gkw282 (2016). doi:10.1093/nar/gkw282

55. Nome, T. et al. High frequency of fusion transcripts involving TCF7L2 in colorectal cancer: Novel fusion partner and splice variants. PLoS One 9, 1–8 (2014).

56. Clark, J. et al. Diversity of TMPRSS2-ERG fusion transcripts in the human prostate. Oncogene 26, 2667–73 (2007).

57. Kimura, K. et al. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. Genome Res. 16, 55–65 (2006).

58. Li, Y., Chien, J., Smith, D. I. & Ma, J. FusionHunter: Identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 27, 1708–1710 (2011).

59. Perner, S. et al. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Res. 66, 8337–8341 (2006).

60. Siegel, R., Miller, K. & Jemal, A. Cancer statistics, 2015 . CA Cancer J Clin 65, 29 (2015).

61. Vogelstein, B. et al. Cancer Genome Landscapes. Science (80-.). 339, 1546–1558 (2013).

62. Wang, K. et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 38, 1–14 (2010).