Title
Clustering of mRNA-Seq Data for Detection of Alternative Splicing Patterns

Author
Johnson, Marla

Publication Date
2017

Peer reviewed|Thesis/dissertation

Clustering of mRNA-Seq Data for Detection of Alternative Splicing Patterns

by

Marla Kay Johnson

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Elizabeth Purdom, Chair
Associate Professor Haiyan Huang
Assistant Professor Nir Yosef

Fall 2017

Clustering of mRNA-Seq Data for Detection of Alternative Splicing Patterns

Abstract

Clustering of mRNA-Seq Data for Detection of Alternative Splicing Patterns

by

Marla Kay Johnson

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Associate Professor Elizabeth Purdom, Chair

Whereas prior methods of studying expression in a cell returned only estimates of gene expression, sequencing of mRNA can provide estimates of the amount of individual isoforms within the cell. As a result, many standard statistical methods commonly used for analyzing gene expression levels need to be modified in order to take advantage of this additional information. Many methods have been developed to study differential isoform expression between known groups but little research has been done utilizing methods of unsupervised learning, such as clustering. One novel question is whether we can find clusters of samples that are distinguishable not by their gene expression but by their isoform usage. That is, instead of using clustering to find groups with shared changes in gene expression, we want to utilize clustering to find groups with shared changes in isoform usage. Here, we propose a novel approach to clustering mRNA-Seq data that identifies such clusters. In order to utilize both gene and isoform information when clustering, we treat the sequencing data as a vector denoting the relative isoform usage of each isoform in a gene. In simulated data, we show that clustering using relative isoform usage values rather than isoform counts is more sensitive to finding clusters based on changes in isoform usage. In a real data set, we demonstrate its performance in finding a technical artifact that resulted in different batches having different isoform usage patterns. Additionally, we also illustrate its usage on several TCGA data sets. Specifically, we looked at whether groups determined from clustering on relative isoform usage were associated with tumor stage or splicing mutations.

To my loves - the big one, the little one, and the littlest one -

my husband Rich and our sons, Max and Charlie

Contents

List of Figures

List of Tables

Acknowledgments

I owe my deepest gratitude to my advisor, Elizabeth Purdom, for all of her support and encouragement throughout my graduate school experience. She has been a wealth of information, ideas, and solace throughout these many years. I am tremendously thankful for her guidance and am so grateful to have had her as my advisor. Additionally, I would like to thank my committee members, Haiyan Huang and Nir Yosef, for their helpful comments.

I would like to acknowledge the members of the Purdom, Sandrine Dudoit, and Terry Speed research groups who provided insight and encouragement over the years. In particular, I would like to acknowledge Anne Biton who performed early work on the batch effect discussed in Chapter 4, as well as Davide Risso who suggested 5′/3′ bias as a possible source of the batch effect. Additionally, over many carrot juices, Christine Ho became a dear friend, and I am grateful for all of our serious and silly conversations over these years.

I was often funded by the Department of Statistics who hired me many, many times as a Graduate Student Instructor. As much as I frequently complained about teaching, I had many students who greatly inspired me, and I am happy to have been part of their education. As a GSI, I am grateful for the numerous opportunities to work with Shobhana Murali Stoyanov and appreciated our conversations about teaching, graduate school, and motherhood. Additionally, I would like to thank La Shana Porlaris in Statistics and Sharon Norris in Biostatistics, who both knew the answers to so many questions so quickly. I would especially like to acknowledge Mary Melinn, who was like a second mother to me during these years.

Lastly, I would like to thank my family. I would like to acknowledge my father, Alfred Johnson. He passed away in April of 2015, and I miss him every day. I would like to thank my sons, Max and Charlie Cohen. They have been giggly, dirty, goofy, and loving reminders that graduate school was not the most important thing in my life. I want to thank my husband, Rich Cohen. He has been my rock during this time and has worked tirelessly for our family. I am humbled by his undiminishing support and numerous sacrificies, and I am beyond grateful for everything he has done for us.

Chapter 1

Introduction

Most genes in eukaryotic genomes produce more than one protein. The expression of these proteins, also called isoforms, is highly regulated in cells, with expression showing both tissue and development specific patterns. Until relatively recently, the laboratory tools used to measure expression of proteins were microarrays. However, microarray are unable to discern different isoforms, and as a result, expression studies focused on gene level expression. Recently, technology advancement including messenger RNA (mRNA) sequencing has made it possible to estimate the expression of individual isoforms. As a result, previous methodology to study gene expression must be modified to be applicable to study isoform expression as well.

Clustering analysis, as described in Chapter 2, is an unsupervised learning algorithm which serves to organize observed data into meaningful subgroups. The idea of clustering using gene expression has been common in cancer studies for a number of years (Sorlie et al., 2001; Alizadeh et al., 2000; Perou et al., 2000). Typically, investigators have identified subtypes in different cancers by finding the shared expression of many genes via clustering. However, mRNA sequencing allows us the chance to cluster using isoform expression rather than gene expression, which may be potentially more informative. Here, we are interested in developing methodology to use mRNA sequencing data in clustering analysis.

While this idea of clustering on mRNA sequencing data usage may be novel, we can study techniques from supervised learning to form this methodology. In Chapter 2, we will describe how differential expression at the isoform and exon level has been studied quite extensively in the last several years. Perhaps the most obvious way to perform isoform clustering is to use the methods developed for gene clustering on isoform estimates instead (Trapnell, Hendrickson, et al., 2013; Leng et al., 2013). One disadvantage to this strategy is the loss of information about the underlying gene structure. That is, rather than only looking at whether the expression of a certain isoform has changed, it may also be useful to examine whether the expression of an isoform has changed relative to the other isoforms in the gene. Because of this, we wanted to look at clustering on the relative abundance of transcript isoforms (Gonzàlez-Porta et al., 2012).

However, the problem now with choosing relative isoform usage to describe each gene is that the value of interest is no longer a simple count. Instead, it is a vector of the proportion of relative expression for each individual isoform. Additionally, since each gene may have a variable number of isoforms, the length of each vector is not constant. Because of this, we must generalize this clustering algorithm to be able to take any type of feature, be it a single value or a proportion, as input. In Chapter 3, we described our methodology in detail. We note that the input into many popular clustering algorithms is typically distance matrices. We suggest that the distance matrix input into a clustering algorithm be viewed as the summation of p distance matrices over each of the p features present in the data. No matter what type of feature is being explored, the distance matrix will have the same form for n individuals; that is, a $n \times n$ matrix. We can then view the overall distance matrix as the weighted sum of the distance matrices for each feature. This is highly generalizable as it allows us to examine any type of feature so long as a distance can be defined between features from two individuals. After combining individual distance matrices, the overall distance matrix may be used in any clustering algorithm which takes a distance matrix as input.

Also in Chapter 3, we describe the simulations we developed to test the data. We developed simulations that allowed us to control gene expression values as well as relative frequency proportions. We tested gene, isoform, and proportion based methods on a variety of settings involving changing expression and relative frequency levels. We found that when changes in isoform expression were due to both gene and alternative splicing signals that clustering based on relative isoform proportion was more sensitive than isoform counts alone.

In Chapter 4, we describe one application of our methodology to a real data set. We were able to identify a previously unidentified batch effect in a publicly available data set. Gene expression clustering or isoform expression clustering did not identify the batch effect as accurately as it was identified using relative isoform frequencies. We show that the batch effect was due to 5′/3′ bias, where one of the plates present show greater coverage relative to the 5′-end of each gene than the other plates did.

In Chapter 5, we describe further results with publicly available data sets associating clustering based on relative isoform abundance to clinical covariates. We show instances where clusters found from clustering on relative isoform abundance are associated with individuals with mutations in known splicing genes. Additionally, we show that clusters formed from proportion clustering contain a relatively higher rate of contrasting isoforms than clusters formed from isoform expression alone. That is, the clusters show differential expression of isoforms coming from the same gene where the gene has at least one isoform that is significantly upregulated and one that is significantly downregulated.

Chapter 2

Background

Our work will introduce a novel method of clustering mRNA-sequencing data which clusters this data by utilizing differential alternative splicing signals. In this chapter, we will introduce some of the background necessary for understanding the motivation of this method. This will include background describing alternative splicing and its role in the development of cancer. Additionally, we will discuss the motivation of clustering and present background on some basic clustering ideas. Lastly, we present some ideas for quantifying alternative splicing in mRNA-sequencing data.

2.1 Alternative Splicing

The human genome is thought to contain around 20,000-25,000 genes, while the total number of proteins expressed is estimated to be around 100,000 proteins. In order to achieve this diversity of proteins, single genes encode multiple proteins. The process by which one gene encodes multiple proteins is called alternative splicing. In the human transcriptome, alternative splicing is the rule rather than the exception, with more than 95% of human genes found to undergo splicing, typically in a tissue or development specific manner (Pan et al., 2008; E. T. Wang et al., 2008).

2.1.1 Biology of Alternative Splicing

As part of the process of turning DNA into proteins, DNA is first transcribed into free floating mRNA in the nucleus. Pre-mRNA consists of regions that will be retained in the final mRNA, termed exons, as well as regions that will be removed, termed introns. This process by which portions of the pre-mRNA are excised is called splicing. In complex organisms, including humans and many common model organisms like fruit flies and mice, mRNA may be spliced in multiple ways, producing different mRNA. The methods by which the final mRNA may differ (shown in Figure 2.1) include skipping or inclusion of exons, alternate 5′ or 3′ splice sites for exons, or intron retention. The varying mRNA

produced by these alternative splicing events allow for one gene to be translated into several functionally distinct proteins.

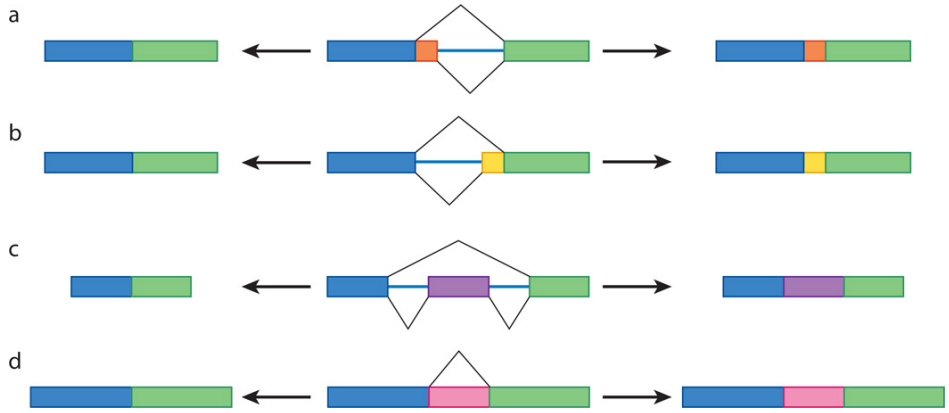


Figure 2.1: **Methods of alternative splicing:** Depicted here are four common methods of alternative splicing, including (a) alternative 5′ splice sites, (b) alternative 3′ splice sites, (c) exon skipping or inclusion, and (d) intron retention. The center column represents the pre-mRNA, while the left and right columns show the mature mRNA after splicing. (Figure reprinted with permission from Nilsen and Graveley (2010))

Splicing is performed by a ribonucleoprotein (RNP) complex known as the spliceosome which is regulated by *cis*-acting regulatory sites and *trans*-acting proteins. The spliceosome is responsible for recognizing the intron/exon boundaries and catalyzing the reactions that remove introns and join together exons. The spliceosome is made up of five small nuclear RNP (snRNP), named U1, U2, U4, U5, and U6, as well as more than 150 proteins. Spliceosome assembly is highly dynamic, being reformed and rearranged on individual introns, allowing for both accuracy and flexibility (Singh and Cooper, 2012).

The major *cis*-acting regulatory sites involved in splicing are splice sites. Splice sites are short, highly conserved sequences that denote the break site between exons and introns. The dinucleotides GT and AG define the 5′ and 3′ boundaries, respectively, of approximately 99% of annotated human introns (Burset, Seledtsov, and Solovyev, 2000). It has been estimated that 15% of the mutations in human disease are found in splice sites (Singh and Cooper, 2012). Other *cis*-acting regulatory sites include the branchpoint and polypyrimidine tract which show weaker sequence conservation than the splice sites (Gao et al., 2008). Disease causing mutations in the branchpoint and polypyrimidine tract are quite rare (Lewandowska, 2013).

Additionally, a large number of splicing regulatory proteins recognize distinct RNA sequences which regulate splicing, a process depicted in Figure 2.2. Splicing silencers are sites where *trans*-acting splicing repressor proteins bind, thereby reducing the likelihood that a nearby site will be used as a splice junction. Alternatively, splicing enhancers act as binding sites for *trans*-acting splicing activator proteins, which increase the likelihood

that a nearby site will be used as a splice junction (Matera and Z. Wang, 2014; Fu and Ares Jr, 2014). Two well-known families of RNA-binding proteins include serine/arginine-rich (SR) proteins, which generally act as splicing enhancers, and heterogeneous nuclear ribonucleoproteins (hnRNP), which generally act as splicing silencers (Fu and Ares Jr, 2014).

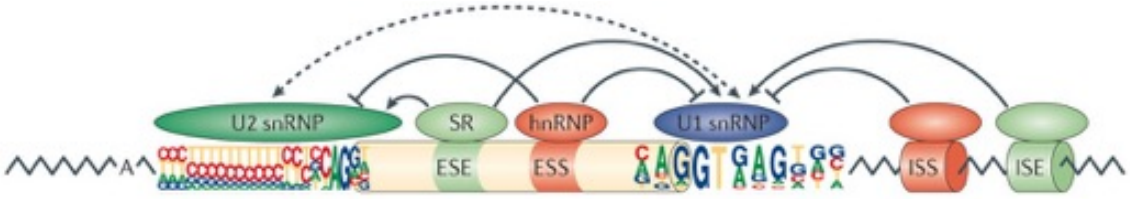


Figure 2.2: **Spliceosome and splicing machinery:** Subunits of the spliceosome (U2 snRNP and U1 snRNP) bind to the 3′ and 5′ splice sites. Binding of the spliceosome is regulated by trans-acting proteins, which bind to splicing enhancers (ESE and ISE) to promote splicing and bind to splicing silencers (ESS and ISS) in order to inhibit splicing. (Figure reprinted with permission from Matera and Z. Wang (2014))

Alternative splicing is a highly regulated process, with over half of alternative splicing events found to be regulated in a tissue specific manner (E. T. Wang et al., 2008). Regulation of expression is thought to be due to differentially expressed splicing factors between different time points or tissue locations. Similarly, another potential mechanism of splicing control may involve regulating the concentration of ubiquitously expressed splicing factors (Chen and Manley, 2009). Regulation must be highly specific as many genes have isoforms that have antagonistic functions. One example of antagonistic isoforms occurs in the gene *VEGF*. One isoform of this gene is used by cancer cells to encourage new vasculature near tumors, while a different isoform is anti-angiogenic and inhibits tumor growth (Qiu et al., 2009).

2.1.2 Alternative Splicing and Cancer

Cancer cells acquire certain properties as they become oncogenic, which include unlimited proliferation of cells, an ability to escape the immune system, promotion of angiogenesis, tissue invasion, evasion of growth suppressors, and immortality (Oltean and Bates, 2014). Many of the genes relevant in these cancer processes are known to have functional diversity as a result of alternative splicing (Sveen et al., 2016). Additionally, abnormal alternative splicing has been shown to be common in cancer (Venables, 2004).

One mechanism which may lead to dysregulated alternative splicing is somatic mutation, which is mutation acquired by a cell. The highly regulated nature of splicing means that mutations in a splice site, the spliceosome, or a regulatory sequence may all potentially affect normal splicing. An example is the gene *SF3B1* which has been found to

have mutations in around 15-20% of uveal melanoma cases and 10% of chronic lymphocytic leukemia tumors (Furney et al., 2013; Gentien et al., 2014; Quesada et al., 2012). Functionally, the SF3B1 protein is involved in the interaction between the U2 snRNP and the branchpoint. Mutations in the *SF3B1* gene are shown to potentially result in the use of cryptic branchpoints, which result in the use of alternative 3'-splice sites. RNA-sequencing identified around 600 instances of alternative 3'-splice site usage that occurred more frequently in patients with *SF3B1* mutations, which generally occurred around 10 to 30 bp upstream of the expected 3'-splice site (DeBoever et al., 2015).

Additionally, many studies have reported cancer specific alternative splicing in the absence of genomic mutations (Venables, 2004). For example, many cancer-associated splicing isoforms are expressed during embryonic development but not in normal adult tissues. This switch may occur due to aberrant regulation of isoforms, allowing for isoforms which promote growth and proliferation to be expressed rather than silenced. In the *VEGF* gene example mentioned earlier, two different isoforms result due to an alternative 3'-splice site in exon 8. The switch between the two isoforms results from different splicing factors promoting the usage of the different 3'-splice site. The pro-angiogenic form of VEGF is promoted by the splicing factors SRSF1 and SRSF5. The anti-angiogenic VEGF isoform is promoted by the splicing factors SRSF6 and SRSF2 (Nowak et al., 2008).

2.1.3 Measuring Isoform Expression

Researchers have used gene expression data in disease research in order to identify differentially expessed genes, build classifiers for diagnosis or treatment, and discover subgroups. Traditionally, these experiments were performed using microarrays. In microarray experiments, mRNA is extracted from tissues or cells and reverse transcribed into cDNA. The cDNA is labeled with a dye and hybridized onto an array. An image is generated of the array and intensity is measured at each location. The measure of intensity will be directly proportional to the level of expression (Tarca, Romero, and Draghici, 2006). However, microarray experiments are limited in their ability to be used to study alternative splicing, though some isoform structure could be discerned though the use of specialized arrays such as exon arrays or exon junction arrays.

Currently, most large studies measure gene expression levels by sequencing of mRNA (Hammerman et al., 2012; Cancer Genome Atlas Research Network, 2013). In mRNA sequencing, mRNA is fragmented and reversed transcribed into short stretches of cDNA. An adaptor is attached to one or both ends and then sequenced (Bu, Chi, and Jin, 2013). The rapid expanse in sequencing technologies has allowed for direct sequencing of mRNA in order to determine the amount of each unique mRNA in a cell. This allows the ability to quantify not just the cumulative amount of expression from a gene region, but also that of individual isoforms within a gene.

The data resulting from microarray experiments is continuous and may be well approximated by a log-normal distribution. Alternatively, RNA-sequencing often results in estimates that are the count of the number of sequences from each gene (depending

on how the estimates are determined). This sequencing data is integer valued and non-negative and does not follow the standard log-normal assumptions of microarray data. Rather, sequencing data may be more appropriately modeled using a discrete count distribution, such as the Poisson or the Negative Binomial distribution (Bullard et al., 2010; Marioni et al., 2008).

Additionally, it is important to note that most current commonly used sequencing technologies still do not allow for the entire mRNA to be sequenced. One exception to this is PacBio Iso-Seq, which performs transcriptome-length sequencing (B. Wang et al., 2016). Rather, the mRNA must be cut into smaller fragments that are then sequenced. This means that estimates of the amount of individual isoforms are not the simple result of counting how many sequences came from particular isoform, but must be indirectly estimated via deconvolution methods (Denoeud et al., 2008; H. Jiang and Wong, 2009; Trapnell, Williams, et al., 2010; Richard et al., 2010; Salzman, H. Jiang, and Wong, 2010; Katz et al., 2010). In genes with multiple isoforms, reads mapping to common constitutive exons must be allocated in a way so as to be consistent with each isoforms expression level. The difficulty in assigning and estimating isoforms with common regions introduces uncertainty in isoform expression estimates (Leng et al., 2013).

Further, moving from a gene-level analysis to a transcript level analysis also introduces an increased number of features to the data. For the purposes of reducing the data size and minimizing noisy data, filtering or feature selection are generally be performed on the complete transcript set.

2.2 Clustering

Cancer studies have often relied on clustering to detect subtypes of tumors based on the expression patterns of genes (see for example (Perou et al., 2000; Sorlie et al., 2001)). These subtypes can have important clinical properties that correspond to disease progression or drug treatment outcomes, creating an important link between the biological mechanisms of tumor cells with the phenotypes observed in tumor patients (Shen et al., 2016).

Mutation is just one way in which the function of a gene can be dysregulated in tumors. Similar phenotypes in tumors can be the result of abnormalities other than mutations, such as a disrupted pathway or shared regulation. This suggests that unsupervised clustering techniques, which do not rely on identifying the source of the abnormality, could provide greater ability to detect dysregulated splicing in tumors. Since mutations in the spliceosome have typically been found at a low prevalence in tumors, such as the gene *U2AF1* which is found mutated at a prevalence of around 5% in acute myleoid lymphoma or lung adenocarcinoma, it may be difficult to correctly identify a cluster containing a relatively small number of the samples unless other causes resulted in similar phenotypes in additional samples (Cancer Genome Atlas Research Network, 2014; Cancer Genome Atlas Research Network, 2013).

Briefly described here are several methods that have been used for clustering gene expression data. For a more comprehensive treatment of cluster analysis, see Everitt et al. (2011) or Aggarwal and Reddy (2013).

2.2.1 Hierarchical Clustering

Hierarchical methods have been popular as a clustering technique for gene expression data. One review noted that, at the time it was published, approximately 95% of studies performing gene clustering used hierarchical methods (Souto et al., 2008). Some early work using hierarchical clustering included subtyping breast cancer (Sorlie et al., 2001) as well as subtyping diffuse large B-cell lymphoma (DLBCL) (Alizadeh et al., 2000). Based on gene expression clustering in the DLBCL study, two distinct expression subtypes were determined. These subtypes were found to have significantly different survival, showing that expression groups could be used to understand disease prognosis (Alizadeh et al., 2000). Hierarchical methods do not produce a single cluster, rather these methods produce a series of partitions, ranging from every sample being in one cluster to every sample being in its own cluster. Typically, this is performed by agglomerative clustering, which begins with every sample by itself in its own cluster, and at each stage, joins together one pair of clusters. After a hierarchical tree of clusterings (also called a dendrogram) is produced, cluster assignments may be determined by cutting the tree at different heights. The dendrogram and heatmap from the DLBCL study can be seen in Figure 2.3.

The parameters used in hierarchical clustering include the choice of distance measure as well as the choice of linkage method. These two parameters are not always independent as the choice of linkage may limit the choice of distance measure used. The linkage method is the distance that is optimized in order to determine which clusters are joined. Some frequently used linkages are described below.

- *Single Linkage* The distance between two groups is the minimum distance between two members.
- *Complete Linkage* The distance between two groups is the maximum distance between two members.
- *Group average or unweighted pair-group average method (UPGMA)* The distance between groups is the average distance between each member of the groups.
- *Centroid or unweighted pair-group centroid method (UPGMC)* The distance between two groups is the distance between their centroids (center of gravity or vector average). This method assumes square Euclidean distance.
- *Simple average or weighted pair-group average method (WPGMA)* The distance between two groups is the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

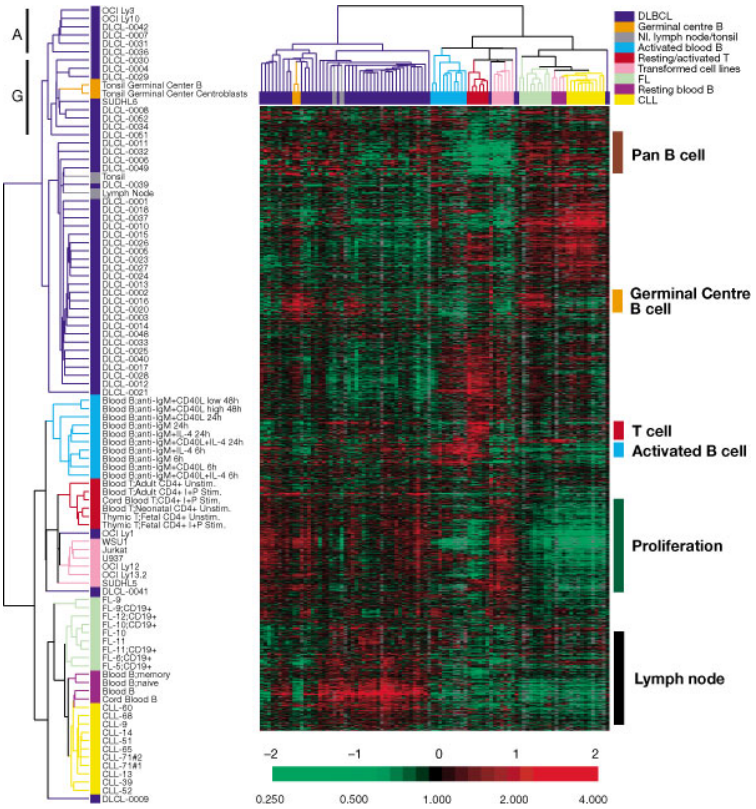


Figure 2.3: **Example of hierarchical clustering:** This is an example of a heatmap ordered by hierarchical clustering. The samples colored in the dendrogram in dark blue are DLBCL samples. Two distinct blocks of expression patterns can be seen, particularly in the lower half the heatmap. (Figure reprinted with permission from Alizadeh et al. (2000))

- *Median or weighted pair-group centroid method (WPGMC)* The distance between two groups is the weighted distance between their centroids, with a weight proportional to the number of members of each group. This method assumes square Euclidean distance.
- *Ward's method* Groups are formed such that the pooled within-group sums of squares is minimized.

An advantage of hierarchical methods is that dendrograms lend themselves to easily interpretable visualizations. Also, due to the fact that a complete hierarchy of clusters is returned, the number of clusters can be chosen after examination of this dendrogram. However, this method becomes computationally intensive as the number of items to be clustered increases. Additionally, some of the linkage methods may not perform well depending on the size of the real clusters, the shape of the data, and the presence of outliers. Since hierarchical clustering is performed stage-wise, the optimal clustering is

found for each stage. That is, a global optimization is not found in hierarchical clustering. For many hierarchical linkages, once a cluster is formed, this cluster may not be broken in any subsequent step.

2.2.2 K-means Clustering

K-means or K-medoids is a partitioning clustering method which separates groups into a set number of disjoint clusters decided *a priori*.

Consider clustering n observations $X_1, ..., X_n$ into m clusters, denoted j. The clustering begins with an initial partitioning of the data by randomly selecting cluster centers, denoted $C_1^0, ..., C_m^0$. A solution is solved iteratively, by repeating the following two steps.

1. *Assignment Step* At step t, assign each object to the cluster whose mean minimizes the within-cluster sums of squares (WCSS).

$$j_t(X_i) = \arg\min_j ||X_i - C_j^t||^2$$

2. *Update Step* Denote the X_i values in cluster k as $S_k^{t+1} = \{X_i : j_{t+1}(X_i) = k\}$. Calculate the new means to be the centers of the observations in the new clusters.

$$C_k^{t+1} = \frac{1}{|S_k^{t+1}|} \sum_{X_i \in S_k^{t+1}} X_i$$

These steps are repeated until a stopping criterion is reached.

Since K-means clustering is a non-convex optimization problem, the solution found in K-means clustering will be a local minima, though not necessarily a global minima. One disadvantage of K-means clustering is that it not always clear what the best K is for clustering, particularly since this is decided *a priori*. Additionally, the solution derived from K-means clustering is heavily dependent on the initial cluster centers. One proposed solution to this is to choose several random initial values, perform K-means clustering on all of those, and choose the clustering that minimizes the objective function as the final solution.

2.2.3 Mixture Models

In mixture model clustering, the data are assumed to be generated by several parametrized distributions, typically Gaussian, and points are assigned to each cluster based on the probability they were generated from that distribution.

The density of a finite parametric mixture model can be written

$$f(\mathbf{x} \mid \pi_1, ..., \pi_G, \theta_1, ..., \theta_G) = \sum_{g=1}^{G} \pi_g r(\mathbf{x}|\theta_g)$$

where π_g is the probability of membership in each subpopulation so that $\pi_g \in [0, 1]$ and $\sum_{g=1}^{G} \pi_g = 1$. The most commonly used distribution is the Gaussian, and the density may be written as

$$f(\mathbf{x} \mid \pi_1, ..., \pi_G, \mu_1, ..., \mu_g, \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_g) = \sum_{g=1}^{G} \pi_g r(\mathbf{x}|\mu_g, \mathbf{\Sigma}_g)$$

Clustering with Gaussian mixture models (GMM) is somewhat related to K-means clustering, but offers the advantage that the covariance structure in GMM may account for correlation between expression values. Typically, cluster assignments are found using the Expectation-Maximization (EM) algorithm so that the distribution parameters and the cluster assignments are iteratively determined with one another (Qu and Xu, 2004; McNicholas and Murphy, 2010).

2.2.4 Self Organizing Maps

Like hierarchical clustering, Self Organizing Maps (SOMs) are useful for visualizing the data and exploratory data analysis. The purpose of SOMs is to develop a map of nodes which are associated with subgroups or patterns in the data set. An example of building a SOM is seen in Figure 2.4. SOMs do not impart as rigid a structure on the data as hierarchical clustering does, although the number of nodes in a SOM is set *a priori*, as in K-means clustering.

The initial step in constructing a SOM is choosing a geometry of nodes. In Figure 2.4, a map of 6 nodes with a 3×2 grid was chosen at the initial step. Additionally, a measure of distance is also needed to build SOMs. The nodes of the SOM are adjusted by iterating through data points, P. The amount a node is adjusted decreases with both increasing distance from P as well as the number of iterations through the data set. That is, the nodes will be adjusted the most in the earliest iterations through the data by points which fall closer to the nodes.

These iterations may continue for 20,000-50,000 times. The point P used at each iteration is determined randomly at the beginning of building the SOM and may be recycled as needed. The final location of the nodes impose structure on the data, with neighboring nodes tending to define related clusters. Alternative structures can be imposed on the data through different initial geometry and different number of nodes, thereby creating a different SOM (Tamayo et al., 1999).

2.2.5 Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) has become a well known method of clustering genes due in part to its use by the Cancer Genome Atlas Project (TCGA) (see for example (Cancer Genome Atlas Research Network, 2013; Cancer Genome Atlas Research Network, 2014)). NMF is an example of dimensionality reduction and decomposes thou-

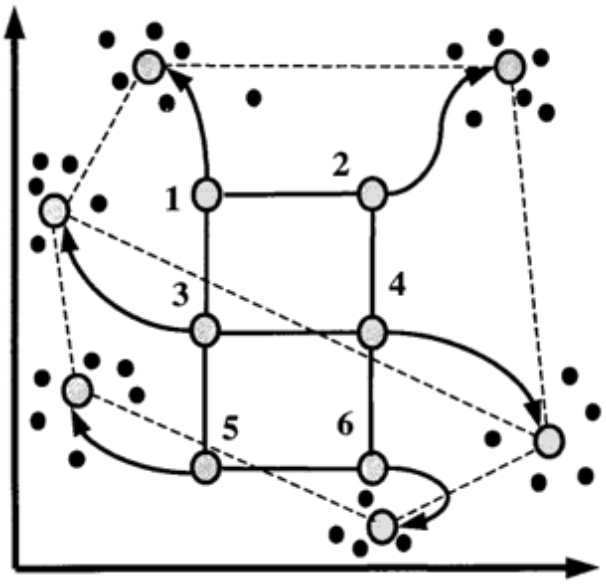


Figure 2.4: **Example of SOM:** In this toy example of SOM, the initial geometry of nodes, represented by large circles, is shown as a 3×2 rectangular grid. During successive iterations of SOM, the nodes move to fit the data. In this figure, this movement is depicted by the arrowed lines. (Figure reprinted from Tamayo et al. (1999), Copyright 1999 National Academy of Sciences.)

sands of genes down to a few metagenes. Samples can then be analyzed by summarizing their gene expression patterns in terms of these metagenes.

Specifically, since mRNA-seq data is expressed as counts, the data can be represented by a n-sample by p-gene (or p-isoform) matrix with non-negative entries. The goal of NMF is to factor this matrix, A, into the product of two matrices which both have positive entries, $A \sim WH$. W is a matrix of size $p \times k$ with each of the k columns representing a metagene. In matrix W, entry w_{ij} is the coefficient for gene i in metagene j. Matrix H is size $k \times n$ with each of the n columns representing the metagene expression pattern of the corresponding sample. That is, for matrix H, entry h_{ij} represents the expression level of metagene i in sample j. After performing this factorization, the H matrix can be used to group the n samples into k clusters. Each sample is clustered according to the most highly expressed metagene in the sample. For example, sample j is placed in cluster i if the h_{ij} is the largest entry in column j (Brunet et al., 2004).

2.2.6 Spectral Clustering

In spectral clustering, gene expression data is reformatted as a graph. In this graph, each gene is a vertex, v_i, and the edges of the graph are weighted by the similarity between the two genes. An affinity matrix can be determined for the entire gene network, and by

using this framework, the clustering problem becomes a graph cut problem and can be solved using spectral graph theory (Huang et al., 2012).

Specifically, if $X_1, ..., X_n$ are the n samples to be considered, a $n \times n$ affinity matrix W can be defined where the entry w_{ij} represents the weight of the edge connecting i and j. The degree of a vertex can be found by summing these weights $d_i = \sum_{j=1}^{n} w_{ij}$. The degree matrix D is the matrix with the degrees $d_1, ..., d_n$ on the diagonal. The unnormalized graph Laplacian matrix can be defined as $L = DW$. In order to form k clusters, the first k eigenvectors, $u_1, ...u_k$ of L are determined. Matrix U is the matrix with these as columns, and y_i is the vector corresponding the the ith row of U. The points $y_1, ..., y_n$ may be clustered by the K-means algorithm into clusters (Luxburg, 2007). Other algorithms to perform spectral clustering are described in Shi and Malik (2000) and Ng, Jordan, and Weiss (2001).

2.2.7 Consensus Clustering

In order to obtain stable clusters, we additionally utilized consensus clustering. The idea in consensus clustering is to find reliable clusters which occur over multiple runs while utilizing a random restart for each run. Consensus clustering assumes that each subpopulation is representative of the data, and as such, clusters formed from each subpopulation should be similar (Monti et al., 2003).

Consider a case in which a sample of the dataset is randomly selected h times. For each resampled dataset, a connectivity matrix, M, is produced in which the entry (i, j) of M is 1 if element i and element j cluster together in the resampled data and 0, otherwise. Additionally, an indicator matrix, I, is also produced where entry (i, j) of I is 1 if the element i and the element j were both present in the resampled data set and 0, otherwise. The consensus matrix can be calculated by taking the average over the connectivity matrices of each resampled dataset, that is

$$\mathcal{M}(i, j) = \frac{\sum_h M^{(h)}(i, j)}{\sum_h I^{(h)}(i, j)}$$

.

That is, for each pair of items, the consensus matrix stores the fraction of time the items are clustered together. A perfect consensus matrix would have only entries of 0 or 1, denoting that items either never or always clustered together. Additionally, the matrix $1 - \mathcal{M}$ can be used as a distance matrix in an agglomerative hierarchical tree construction algorithm to yield a dendogram of item adjacencies. This final clustering on the $1 - \mathcal{M}$ represents the final cluster assignments (Monti et al., 2003).

In addition to determining the most stable cluster assignments, we also found empirically that consensus clustering was useful in identifying possible outlier samples. Outliers were determined by tracking which elements tended to cluster with only themselves or a few other samples.

2.3 Clustering mRNA Sequencing data

Clustering analysis serves to organize observed data into meaningful subgroups in an unsupervised manner. Clustering has been a common technique in microarray based gene expression studies and using clustering as part of isoform expression studies will no doubt prove quite useful as the expression of individual isoforms is known to be highly regulated. However, due to the additional complexities of sequencing data discussed here, including the choice of an appropriate distribution, the increased uncertainty in the estimates, and the increased size of the data, the methods used for microarray clustering may not applicable or optimized for use with sequencing data.

Research done thus far on clustering of mRNA sequencing data has been performed only on gene expression estimates, though the methods developed may potentially be applied to isoform expression estimates as well. A related area of research is differential alternative splicing of isoforms. A major difference between this and clustering is that clustering is an unsupervised method, while differential isoform expression or differential alternative splicing analysis is performed on known groups. However, many ideas of quantifying isoforms or splicing may be applicable to the problem of clustering isoforms. In this section, we will describe some techniques used in differential expression to quantify the effects of splicing that may potentially also be useful in clustering analysis.

2.3.1 Clustering Gene Expression

For microarray data, the square Euclidean distance is often the choice for the dissimilarity measure. This choice is appropriate as the log likelihood ratio statistic for testing the equality of two means under a Gaussian model of the data is equivalent to the square Euclidean distance. However, as previously mentioned, mRNA sequencing expression data is count based so assuming a log normal distribution is no longer appropriate. Using discrete distributions, including the Poisson (Bullard et al., 2010; Marioni et al., 2008) and Negative Binomial distributions (Robinson, McCarthy, and Smyth, 2010; Anders, 2010), for studying gene differential expression in gene counts have frequently been used, and it seems a natural extension to apply these same models to gene clustering (Witten, 2011; Si et al., 2014).

In the discrete count based model, X_{ij} indicates the total number of reads mapping to gene j in observation i, while y_i indicates the class of the ith observation. C_k contains the indices of the observations in class k. These models allow for variability in the number of counts per sample (s_i), the number of counts per feature (g_j), and differences in counts due to class membership (d_{kj}). If d_{kj} is greater than 1, the jth feature is over-expressed relative to the baseline in the kth class. If d_{kj} is less than 1, the jth feature is under-expressed relative to the baseline in the kth class. Modeling gene expression using the Poisson distribution can be parameterized as:

$$X_{ij}|y_i = k \sim \text{Poisson}(N_{ij}d_{kj}), N_{ij} = s_i g_j,$$

Modeling gene expression using the Negative Binomial distribution can be parameterized as:

$$X_{ij}|y_i = k \sim \text{Negative Binomial}(N_{ij}d_{kj}, \phi_j), N_{ij} = s_i g_j,$$

As described in (Witten, 2011), in fitting the Poisson model, the maximum likelihood estimate for N_{ij} is $\hat{N}_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{X_{\cdot\cdot}}$. The parameter d_{kj} could be estimated by $\hat{d}_{kj} = \frac{X_{C_k j}}{\sum_{i \in C_k} \hat{N}_{ij}}$.

Witten (2011) used the Poisson model to developed a dissimilarity measure for count data. In this case, the null hypothesis being tested in this model is $d_{ij} = d_{i'j} = 1$, while the alternative hypothesis is that d_{ij} and $d_{i'j}$ are unconstrained estimates. The resulting log likelihood ratio statistic may be used as a measure of dissimilarity between observation i and observation i':

$$\sum_{i}^{p}(\hat{N}_{ij} - \hat{N}_{ij}\hat{d}_{ij} + X_{ij}\hat{d}_{ij} + \hat{N}_{i'j} - \hat{N}_{i'j}\hat{d}_{i'j} + X_{i'j}\hat{d}_{i'j})$$

After performing hierarchical clustering utilizing this distance, Witten (2011) evaluated the performance of Poisson clustering using clustering error rate (CER). In determining CER, the performance of a partitioning Q is measured against the true partitioning P. The indicators $1_{P(i,i')}$ and $1_{Q(i,i')}$ are defined for whether i and i' are in the same group in partitions P and Q. The CER is then defined by

$$\sum_{i>i'} \frac{|1_{P(i,i')} - 1_{Q(i,i')}|}{\binom{n}{2}}.$$

Using this metric, Witten (2011) found that hierarchical clustering using the Poisson based distance produced more accurate clusters than clustering with the square Euclidean distance.

Similarly, Si et al. (2014) clustered mRNA sequencing via mixture models similar to described in Chapter 2.2.3. However, instead of assuming the data was generated from a Gaussian distribution as described in Chapter 2.2.3, the model based methods for mRNA sequencing data assumed the data was generated from either a Poisson or Negative Binomial distribution. Si then used the EM algorithm to iteratively estimate the parameters and perform the clustering. In order to compare clustering strategies, this group examined pairwise sensitivity, which is defined as the proportion of pairs clustered together in the true partitioning P that are also clustered together in Q. Additionally, they also calculated pairwise specificity, which is defined as the proportion of pairs in clustered to different groups in P which are also clustered to different groups in Q. Lastly, they calculated Normalized Mutual Information, a measure of the amount of information one random variable contains about another. Using these metrics, Si et al. (2014) found that Poisson and Negative model-based clustering method outperformed clustering produced from K-means clustering and SOM.

2.3.2 Isoform Expression

Perhaps the most obvious way to perform isoform clustering is simply to use the methods developed for gene clustering on isoform estimates instead. Isoform estimates are similar in structure to that of gene estimates and therefore such clustering could make use of similar existing procedures. However, as already described in this chapter, estimates of isoform expression contain greater uncertainty due to the presence of multiple isoforms in many genes. Some methods used for differential expression analysis have been developed to account for this increased uncertainty due to the presence of isoforms, including CuffDiff2 and EBSeq.

CuffDiff2 first models the variability in fragment count for each isoform across replicates using the Negative Binomial distribution (Trapnell, Hendrickson, et al., 2013). During the initial phase, for sample i and isoform k in class y_i, the data is modeled as:

$$X_{ijk}|y_i \sim \text{Negative Binomial}(p_{jk}, r_{jk})$$

The mean and variance are estimated by fitting a generalized linear model in the replicates in each class. Then fragments are assigned to transcripts using maximum-likelihood estimation. The uncertainty in this estimation due to ambiguously mapped reads is modeled as a Beta distribution (Trapnell, Hendrickson, et al., 2013):

$$p_{jk} \sim \text{Beta } (\alpha_{jk}, \beta_{jk})$$

The resulting mixture is a Beta Negative Binomial distribution:

$$X_{ijk} \sim \text{Beta Negative Binomial}(r_{jk}, \alpha_{jk}, \beta_{jk})$$

Cuffdiff2 estimates gene and transcript level expression, the variance in these expressions, and the covariances between isoforms of the same gene. To test for significance between transcript levels in different classes, CuffDiff2 uses the log-transformed ratio of expression between groups divided by the variance of the transformed ratio. This statistic roughly follows a standard Normal distribution (Trapnell, Hendrickson, et al., 2013).

EBSeq is an empirical Bayes approach which also uses expression estimates from reconstructed isoforms. EBSeq models counts using a Negative Bionomial distribution, while utilizing a Beta prior to model fluctuations in technical and biological variation (Leng et al., 2013). For isoform k of gene j in sample i, the expected count can be parameterized within class y_i as:

$$X_{ijk}^{y_i}|r_{jk}l_i, q_{jk}^{y_i} \sim \text{Negative Binomial}(r_{jk}l_i, q_{jk}^{y_i})$$

Note that l_i represents the total library size in sample i. Additionally, this model assumes a prior distribution on q_{jk}^{C}, which in this case is a Beta distribution. Included in the model is a variable I_j, which is described as a measure of isoform complexity. For example, it may be the number of isoforms from a gene or a value derived from read

alignment, such as an isoform's mappability score or credibility interval. The model is parameterized as:

$$q_{jk}^{y_i}|\alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_j})$$

Estimates of isoform means and variances are obtained via method-of-moments. Estimates for α and β^{I_g} are derived using the EM algorithm. Once the parameters have been estimated, the posterior probability of differential expression may be obtained using Bayes rule (Leng et al., 2013).

2.3.3 Relative Exon Usage

Other methods compare quantities that may be obtained from sequencing data without reconstructing the isoforms, such as exon or intron usage. These methods assume that differences in isoform expression should be also distinguishable in the usage of some individual introns and exons. One commonly used method incorporating this idea is DEXseq (Anders, Reyes, and Huber, 2012). DEXseq, like other count based methods, models the number of reads aligning to a bin (typically an exon) using a Negative Binomial distribution. For gene i in bin l for sample j, the counts are assumed to be distributed as:

$$y_{ijl} \sim NB(s_j \mu_{ijl}, \phi_{il})$$

For each gene, a linear model is fit where the mean of each gene is predicted as

$$\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{ic_j}^C + \beta_{ic_jl}^{EC}.$$

The linear predictor μ_{ijl} is decomposed into four factors. The β_i^G is the baseline expression of gene i while β_{il}^E accounts for the expected fraction of the reads that further overlap with counting bin l. The other two terms account for how counts change under different conditions, denoted c_j. $\beta_{ic_j}^C$ accounts for the difference in gene expression between conditions, while $\beta_{ic_jl}^{EC}$ is the effect on the fraction of reads falling into bin l. A non-zero $\beta_{ic_jl}^{EC}$ indicates that the counting bin ijl was differentially used between different conditions, while a non-zero value of $\beta_{ic_j}^C$ indicates differential expression of the gene (Anders, Reyes, and Huber, 2012). While typically the counting unit used are the exons of each gene, some methods have expanded on DEXSeq by enumerating reads that align to splice junctions or intron retentions (Y. Li et al., 2015; Hartley and Mullikin, 2016).

2.3.4 Percent Spliced In

This method specifically counts reads that show evidence of inclusion or exclusion of an exon (or also potentially an intron). Single-end reads that align to the alternative exon or to the splice junction with an adjacent exon provide support for inclusion of that exon. Reads that align across the splice junction between constitutive exons support the exclusion isoform. The ratio calculated from these read counts is often called percent spliced

in (PSI or ψ) or exon inclusion percentage and is defined by(E. T. Wang et al., 2008; Katz et al., 2010):

$$\psi = \frac{\text{Number of reads including exon}}{\text{Number of reads including exon + Number of reads excluding exon}}.$$

A modification to this the calculation of ψ was proposed by the method MISO. In this method, when paired-end sequencing is performed, information about the library insert length distribution for paired end reads may be used to further assign additional reads to the inclusion or exclusion isoforms and improve the estimation of ψ (Katz et al., 2010).

An obvious difference between this method and the previous methods of quantifying alternative splicing is that the measure is no longer a count but rather a proportion. Due to this, a different model is required to describe the data. In one method, MATS, the joint prior probability of ψ_{ic_1} and ψ_{ic_2} is modeled as a multivariate uniform distribution. The biological motivation for this is that between any two biological conditions, only a small number of exons will undergo differential alternative splicing, while most will show similar splicing patterns. The model of MATS describes the data as

$$(\psi_{ic_1}, \psi_{ic_2}) \sim \text{MultiVarUniform}\left(0, 1, cor = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$
$$y_{ij}|\psi_{ic_j} \sim \text{Bin}(n_{ij}, p_{ic_j})$$

for gene i for sample j under condition c_j.

Our group developed a method modeling this type of data using an overdispersed binomial distribution from the double exponential family (Ruddy, M. Johnson, and Purdom, 2016). This class of distributions adds a dispersion parameter to any member of the exponential family (Efron, 1986). We proposed an empirical Bayes strategy for producing a shrunken estimate of dispersion which effectively detected differential proportional usage.

2.3.5 Relative Abundances of Transcript Isoforms

Our strategy here, however, is on evaluating the *relative* isoform usage within a gene: a measure of the tendency of a gene to prefer one isoform over another. We will consider this framework in more detail in the next chapter, but briefly, in this method, we can consider the data for each gene as a having a multinomial distribution.

$$(y_{ij1}, ..., y_{ijk}) \sim \text{Multinomial}(N_{ij}, p_{ij1}, ..., p_{ijk})$$

for gene i with k isoforms in sample j. This methodology bears some similarity to that of calculating the exon inclusion percentage in that the count for each isoform is normalized by the overall gene count, N_{ij}, so the summary is a vector of proportions rather than in terms of counts.

Gonzàlez-Porta et al. (2012) used a similar way of quantifying data, which they termed a splicing ratio, to study alternative splicing variability. Using a method modified from Anderson (2001), they used the Hellinger distance, which we will discuss in Chapter 3.1.1, to compute difference between two individuals' proportion vectors. For each gene, they calculated the centroid, **c**, the point which minimized the sum of squared distances between itself and each point in the set of sampled points. The variability of each gene was then measured by

$$SS = \sum_{i=1}^{n} d^2(i, \mathbf{c}),$$

where $d^2(i, \mathbf{c})$ is the squared Hellinger distance. Using these metrics, this group stated that splicing ratios remained relatively constant in populations and suggested that 60% of the variability seen in isoform levels is actually due to changes in gene expression rather than changes in splicing ratio.

Chapter 3

Methodology

In many instances when working with mRNA-seq data, the count estimates from different isoforms are combined together to return a measure of gene expression. Beyond calculating the gene expression, the additional information provided by knowing the individual isoform expression is simply discarded. Alternatively, if one clusters instead on isoform counts, we typically lose the information about which gene the isoform is derived. A potential strategy may be trying to incorporate gene and isoform information in one feature. Here, we examine a strategy for describing each gene not simply by a count over all of its isoforms but instead as the relative isoform usage of the isoforms in the gene.

In this chapter, we will discuss in detail our methodology for clustering on isoform usage. We are particularly interested in mRNA-seq data showing differential alternative splicing, and we will examine how clustering on gene counts, isoform counts, and relative isoform usage differ in finding this signal. Through simulation, we show that our method was highly accurate in detecting clusterings due to differential isoform usage.

In addition, we will describe a modification to our methodolgy where we incorporated sparse clustering in hopes of achieving more accurate clustering. Since only a subset of the features actually contain the signal we are interested in, we utilized sparse clustering as a framework to select those features. Ideally, the potential benefit of sparse clustering would be improved cluster accuracy as well as additional information about which features are driving the clustering. However, we found sparse clustering resulted in highly variable and less accurate results compared to standard clustering with variance filtering applied to the data.

3.1 Description of Methodology

Our goal is to cluster n individuals into k clusters. Our underlying motivation is expression data from mRNA-seq experiments. As described in the first chapter, mRNA-seq data is typically summarized as gene expression counts. We can denote the gene expression

for individual i for gene j as x_{ij}. This value is a singular, discrete number. We can utilize many different distances for this calculation, but in many clustering techniques, the distance between two individuals is typically calculated using Euclidean distance. That is, we can find the distance between two individuals i and i' by the equation

$$d_{i,i'} = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

.

The overall distance matrix D is then a $n \times n$ matrix with distance $d_{i,i'}$ in position (i, i'). However, this equation would only apply in the simple case where the features are single-valued entries. In order to have single-valued entries when we work with mRNA-seq data, we either combine all isoform expression estimates into one gene expression estimate, as above, or we ignore the gene grouping structure of isoform estimates and compare all isoforms individually.

An alternative to these methods would be to use the isoform expression estimates in a way that utilizes the underlying gene structure. Looking at the isoforms in the context of their gene structure could be useful as we are privy to information not only about how each individual isoform changes but also about how they change relative to each other. An impediment to grouping the isoform data by gene is that each gene may have a different number of isoforms. We saw in our own data that genes with multiple isoforms could have anywhere from 2 isoforms to over 15 isoforms expressed (Figure 3.1).

Figure 3.2 gives an idea of the increased complexity that arises from maintaining the isoform expression by gene grouping. In a typical two-dimensional $n \times p$ data set, the value (i, j) is an expression estimate for individual i in gene j. In this new framework, the value (i, j) is no longer a single number, but rather, a vector of values $p_{ij} \in R^{K_j}$, where K_j is the number of features or isoforms in gene j, varying for each j.

It is clear data in this structure cannot be used in the calculations of the overall distance matrix we have already discussed. However, we note that the overall distance matrix may be decomposed slightly differently than how it was originally presented. Since the distance between two individuals is the overall sum of the distance between each of the p features for those two individuals, we could instead calculate p distance matrices, one for each feature, and then find an overall distance object by summing up all p distances matrices. In our continuing example, for feature j, the distance between individual i and individual i' would be

$$d_{i,i',j} = (x_{ij} - x_{i'j})^2$$

We can see that it is still true that

$$d_{i,i'} = \sum_{j=1}^{p} d_{i,i',j}$$

For each feature j, the distance matrix D_j is a $n \times n$ matrix with distance $d_{i,i',j}$ in position (i, i'). Since the overall distance matrix compares distances between individuals

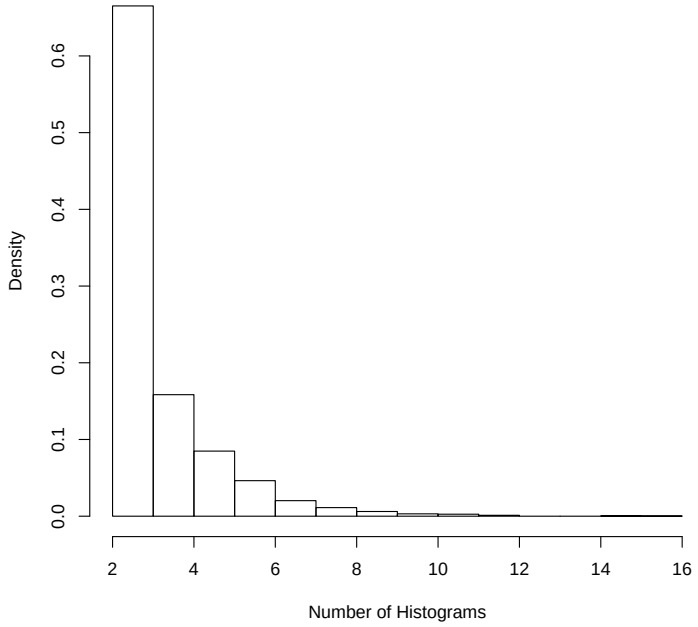


Figure 3.1: **Histogram of number of isoforms per gene:** This histogram shows the distribution of the number of isoforms present in genes with multiple isoforms in the TCGA acute myeloid leukemia data set. The maximum number of isoforms in this case was 16 isoforms, with a median of 3 isoforms and a mean of 3.3 isoforms

at the pairwise level, the distance matrix is always size $n \times n$, regardless the structure of the feature. As long as the per feature distance between two individuals i and i' is described by a single value, we can think of the overall distance matrix as the sum of the distance matrices of each feature j, denoted D_j.

$$D = \sum_{j=1}^{p} D_j$$

By framing the calculation of the overall distance matrix as the sum of individual distance matrices, we are able to utilize more complicated features than ones that are represented by a single value, as is the case for gene counts or expression data. This creates a flexible clustering strategy that allows the feature information to contribute to the clustering only in the context of its relationship to other features in the group. Additionally, as will be formalized in our discussion of sparse clustering in Section 3.4, we could potentially incorporate different weights for each feature, denoted w_j, allowing the overall

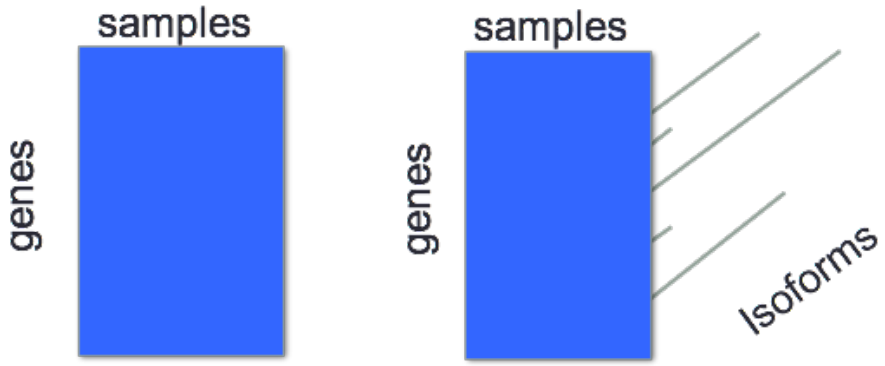


Figure 3.2: **Depiction of the data used in gene and isoform count clustering:** The data structure on the left is a gene count matrix with n observations and p genes which has the gene count in position (i,j). The data structure on the right is composed of individual isoform counts, which are only considered relative to other isoforms from the same gene. This $n \times p$ matrix will have a vector in position (i,j) that describes the relationship of all isoforms in gene j.

distance matrix to be a weighted sum of the per feature distance matrices.

$$D = \sum_{i=1}^{p} w_j D_j$$

Since we are no longer summarizing the isoform expression estimates as one gene expression estimate, our data is no longer a single value x_{ij} but rather a vector $x_{ij1}, x_{ij2}, ..., x_{ijK_j}$. We could use this vector to describe each gene, though it seems advantageous to instead normalize this vector by the gene expression estimate. Therefore, the feature we use to describe each gene is the relative isoform frequency of the gene:

$$p_{ij} = \left(\frac{x_{ij1}}{x_{ij}}, \frac{x_{ij2}}{x_{ij}}, ..., \frac{x_{ijK_j}}{x_{ij}} \right)$$

While our initial motivation was mRNA-seq data, the question of clustering relative isoform usage can be stated more generally as the problem of clustering data when the features, in this case the isoform estimates, are known to belong to a predefined group, in this case the gene. Gene estimates are then a summary statistic (the sum) of the features in the group. We can then think of gene expression clustering as clustering of a summary statistic of the group members, while isoform clustering is clustering of the individual features while ignoring the group membership.

Describing the data in terms of distance matrices is useful in that many clustering methods require a distance matrix as input. For example, K-means clustering attempts to partition the clusters by minimizing the within cluster sum of squares. We can write the

within cluster sum of squares in terms of the distance function:

$$\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} \sum_{j} d_{i,i',j}.$$

We note that $d_{i,i,j}$ may denote any dissimilarity measure so long as a distance for feature j is defined between i and i'. We see by recasting K-means clustering in terms of distance matrices that we have generalized the results. That is, regardless of the type of feature used, we may used K-means clustering so long as a distance $d_{i,i,j}$ may be defined for this feature.

3.1.1 Choice of Distances

Prior to performing clustering on count based methods, we first processed the data by taking the log of the gene and isoform counts. We then found the pairwise distance within each feature by utilizing the Euclidean distance as the dissimilarity measure between observations.

For the relative isoform usage method, we also needed to find a distance for each gene, though it is less obvious in this case which distance to use. While there are numerous distances defined between vectors that lie on the simplex (see M. M. Deza and E. Deza, 2014, for a review), we considered some of the most popular: χ^2 distance, Euclidean distance, Jeffrey's divergence, and Hellinger's distance. Another group analyzing relative isoform usage utilized Hellinger's distance in their analysis (Monlong et al., 2014; Gonzàlez-Porta et al., 2012; Ferreira et al., 2014). Dropping the dependence on the gene j, the distance between two proportions, p_i and $p_{i'}$, is defined as:

- Squared χ^2-measure:

$$d(p_i, p_{i'}) = \sum_{k=1}^{K} \frac{(p_{ik} - p_{i'k})^2}{p_{ik} + p_{i'k}} \tag{3.1}$$

- Euclidean distance:

$$d(p_i, p_{i'}) = \sqrt{\sum_{k=1}^{K} (p_{ik} - p_{i'k})^2} \tag{3.2}$$

- Jeffrey's divergence:

$$d(p_i, p_{i'}) = \sum_{k=1}^{K} (p_{ik} - p_{i'k}) \ln \frac{p_{ik}}{p_{i'k}} \tag{3.3}$$

- Hellinger distance:

$$d(p_i, p_{i'}) = \sqrt{2 \sum_{k=1}^{K} (\sqrt{p_{ik}} - \sqrt{p_{i'k}})^2} \tag{3.4}$$

We also evaluated a distance based on the multinomial log-likelihood ratio test, similar to the Poisson-based distance described in Chapter 2.3. This distance measure differs from the other ones we explored as this is calculated on isoform counts directly rather than isoform relative frequency. In the multinomial log-likelihood test, calculations of this value are used to establish whether isoform usage among different individuals have the same distribution. Values of the multinomial log-likelihood ratio are larger when it is not likely that the counts have been drawn from the same multinomial distribution, a feature we will exploit as a distance.

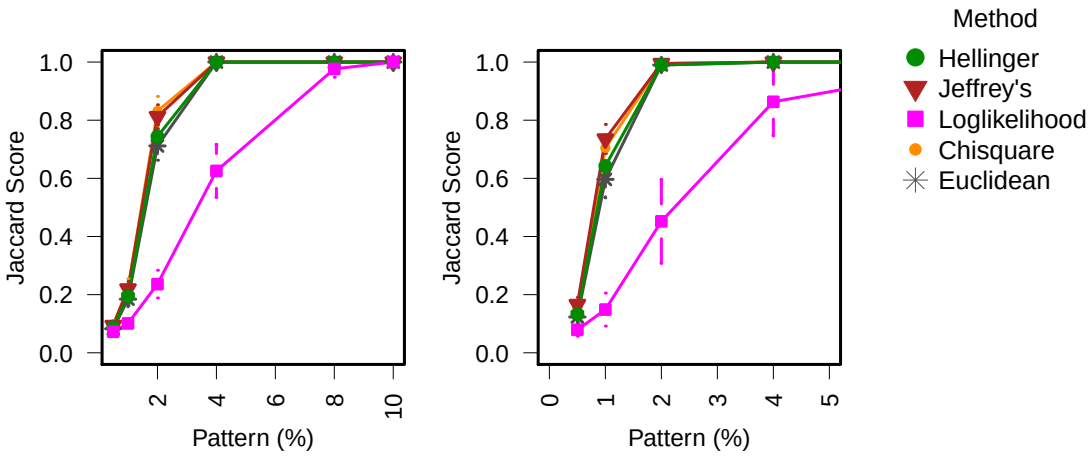
- Log-likelihood based distance (Berninger et al., 2008; Witten, 2011):

$$d(i,i') = \frac{\mathcal{L}(x_i, x_{i'} | p_i \neq p_{i'})}{\mathcal{L}(x_i, x_{i'} | p_i = p_{i'})} = \sum_{k=1}^{K} x_{ik} \ln \frac{x_{ik}}{x_{i+}} + x_{i'k} \ln \frac{x_{i'k}}{x_{i'+}} - (x_{ik} + x_{i'k}) \ln \frac{x_{ik} + x_{i'k}}{x_{i+} + x_{i'+}} \tag{3.5}$$

3.1.2 Implementation of Clustering

After calculating these per-feature distances, we then take a weighted sum of these per-feature distance matrices to calculate one distance object. A distance object itself may be used as input into many different clustering algorithms. Here, we utilized hierarchical clustering from the package **stats** and K-medoids clustering functions from the package **cluster** in R, which both take distance objects as input. For clustering methods that take raw data rather than distance matrices, it is possible to perform multidimensional scaling (MDS) on the distance object to return a two-dimensional data matrix which preserves the distances between individuals. Many clustering algorithms which do not take distance objects as inputs will instead take this form of data, such as the K-means clustering function from the package **stats**.

We examined how the different distance measures behaved on the same data set. We were interested in whether the distance measures showed much difference in their behavior. In order to quantify clustering assignments, we used a measure of similarity, or accuracy, called the Jaccard index (discussed further in Section3.2.3). In order to measure performance, we examined the accuracy of the cluster assignments over different effect sizes, the percentage of genes which contained a known pattern. We noticed that the Euclidean and χ^2 distances resembled Hellinger and Jeffrey's divergence, respectively, in our simulations (see Figure 3.3) and on implementation on real data. Due to this, we only show the Hellinger distance, Jeffrey's divergence, and Log-likelihood distances in subsequent plots.



(a) Only proportion differences (b) Gene and proportion differences

Figure 3.3: **Comparison of different dissimilarity measure:** Plotted are Jaccard Scores (y-axis) versus the percent of genes with the variable clustering pattern (x-axis). Different lines correspond to different methods of measuring distance between proportions, indicated by the legend. In this simulation, the relative isoform usage is different across the nine clustering groups, but gene expression remains the same for all genes (Case 2, Figure 3.5b).

3.2 Details of Simulations

In developing simulations, we wanted to address two particular concerns. First, we wanted to show that the results obtained from the relative isoform usage method were accurate in detecting a differential alternative splicing signal. Secondly, we wanted to understand whether the obtained results were novel and useful or if the results were simply redundant from clustering on gene or isoform counts. To examine these concerns, we developed simulations meant to mimic mRNA-seq results in real data that contained differentially alternative spliced signals. We clustered the data based on gene counts only, isoform counts only, and relative isoform usage.

One concern we had was that genes may typically contain one major isoform showing high expression with other minor isoforms showing low expression. In real data (Figure 3.4), we saw high correlation between gene expression and the isoform with the highest expression. Across all genes with multiple isoforms, the median of the correlation coefficient between gene expression and major isoform expression was 0.93. Since it tends to be the case that gene expression and expression of the major isofrom is highly correlated, clustering on isoform counts should return similar results to clustering on genes counts, with the major isoforms behaving like the gene and the minor isoforms adding noise. This suggests clustering on isoform counts would be adding unnecessary complexity in order to simply return the same results as clustering on gene counts.

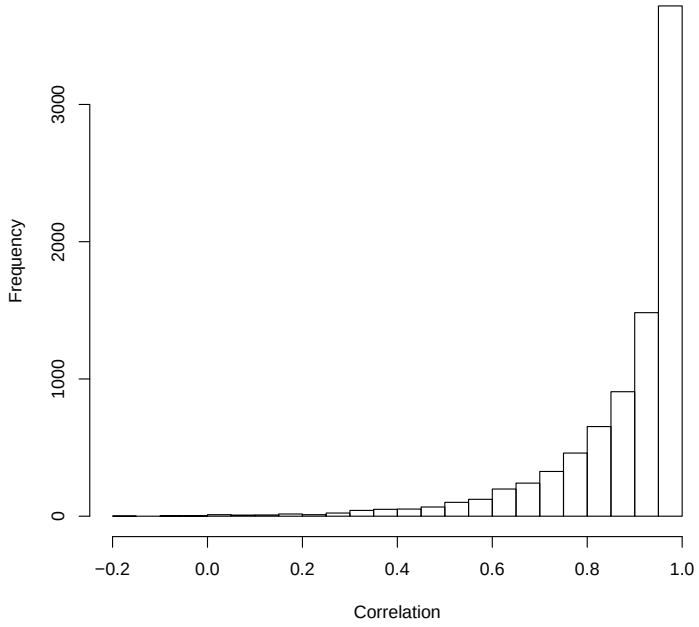


Figure 3.4: **Correlation between gene counts and major isoform counts:** Using isoform counts from the TCGA acute myeloid leukemia data set, this histogram shows the distribution of the correlation coefficients between the gene counts and the counts of the major isoform in each gene. The median of these correlation coefficients was 0.93.

Additionally, clustering on isoform counts is attempting to find a different signal than clustering on relative isoform usage. In the first case, we are trying to find groups where any isoform counts vary between the groups. However, in the second case, we are trying to find groups where the relative usage varies between groups. That is, the isoform expression varies relative to other isoforms in the gene. As an example, in the first case, we may find two clusters where some isoforms double in the expression between the two groups. However, if those isoforms are from the same gene so that all the isoforms in the gene double their expression, the gene would show no signal of differential alternative splicing although it shows a signal of differential isoform expression. In real data, this could be potentially the case for isoforms which share the same promoter. This may also happen in cases where genes have multiple isoforms that are not easily deconvolved. That is, isoforms may all encode many of the same regions, which does not allow for unambiguous assignment of reads. In such cases, reads that align to the ambiguous portions of isoforms are proportionally allocated based on the number of reads that align to the portions that are unique.

In these simulations, we did not perform any additional filtering based on expression levels. In real data, however, this was a reasonable step in order to reduce the size of the data set. Additionally, genes with very low expression will not return meaningful estimates of isoform usage and should be discarded.

3.2.1 Clustering Scenarios

We examined several different ways in which clusters with different proportional usage of isoforms could occur. For any particular gene, we considered that gene expression and relative isoform frequency could interact in three ways.

1. Gene expression counts differ between groups, while the proportion levels are constant across groups (Figure 3.5a).

2. Gene expression counts are constant across groups, while the proportion levels vary between groups (Figure 3.5b).

3. Both the gene expression and proportion levels vary across groups (Figures 3.5c and 3.5d).

We also simulated a setting where some genes followed Case 1 (i.e. had only gene expression differences) while a smaller fraction of genes followed Case 3 (both gene and proportion differences in the same gene). We expected this to be the most likely scenario biologically. That is, many genes show differential gene expression, while only a few genes some differential isoform usage.

In Case 1, we expect clustering on the gene counts and isoform counts to be effective and do not expect useful results from clustering on relative isoform frequency. In Case 2, we expect clustering on isoform counts and relative isoform frequency to be useful but do not expect useful information from clustering on gene counts.

In simulations where both gene and proportion clustering coexisted (Case 3), we need to be able to distinguish between the effect of the gene clusters and the proportion clusters. We set the isoform frequency groups to be different (and smaller) than that of the gene expression levels. The proportion clusters were either nested within the gene groups or spanned across gene groups.

If proportion clusters were nested, the relative isoform frequency within each gene group shifted, while overall gene expression stayed the same. We first simulated three different gene groups, and within each gene group, we further simulated three isoform frequency groups, for a total of nine groups showing different relative isoform frequencies. This combination also leads to nine groups showing different isoform expression levels (Figure 3.5c).

In the case where the proportion clusters are not all nested within gene clusters, we again begin with three groups with different gene expressions and simulated six groups

showing relative isoform frequency shifts. Unlike the previous case, these different subgroups were no longer contained in just one gene level group. This is important to the specific case of clustering on isoform levels since changes in gene levels as well as changes in relative isoform usage will alter isoform values. If relative isoform usage is constant but gene levels vary, we would see isoform levels change relative to the change in gene level. If gene levels are constant but relative isoform usage varies, we would see isoform levels shift according to the new proportion levels. If both gene levels and relative isoform usage vary, the isoform levels would vary based on the combination of shifts in both gene levels and proportion usage. Theoretically, clustering directly on isoform expression levels should be able to detect both gene and proportion differences. The distribution of the six different relative isoform frequency groups across three gene expression groups resulted in nine clusters with different isoform expression values (Figure 3.5d). In our plots of this case, we show the ability of isoform clustering to detect the three gene clusters as well as the nine isoform clusters.

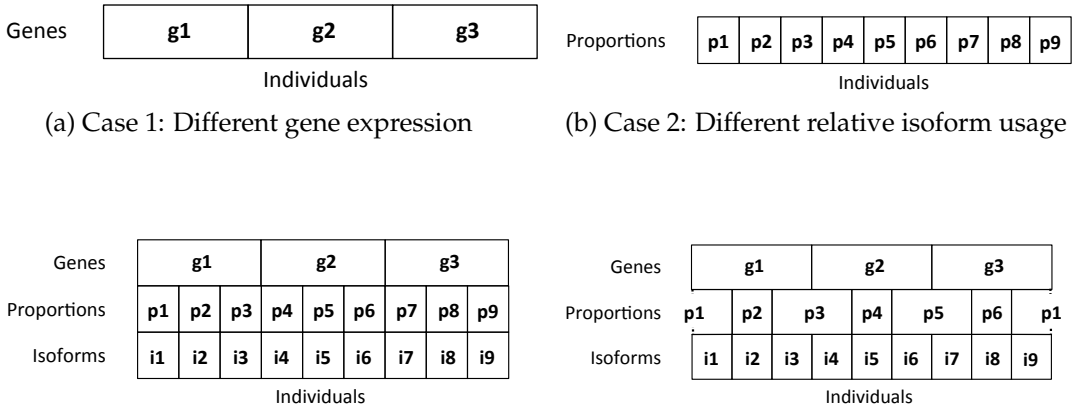
To simulate each of these cases independently, we simulated a fraction of all the genes to contain a set pattern in their gene expression with isoform proportions corresponding to one of the cases above, while the remaining genes had both constant gene and isoform expression (no clustering signal). The percentage of genes with the variable relative isoform usage pattern varied from 0.5%, 1%, 2%, 4%, 8%, and 10%.

In the more complex setting where some genes followed Case 1 while other genes followed Case 3, we held the percentage of genes following Case 1 fixed at 25% and allowed the percentage of genes following Case 3 to vary according to the same proportions given above (again with the remaining genes held constant in gene and isoform expression).

3.2.2 Generating Isoform Counts

For each simulation, we simulated 5,000 genes across 135 samples. Prior to performing the simulations, exploration of the TCGA acute myeloid leukemia (LAML) data set (Cancer Genome Atlas Research Network, 2013) had suggested that the median and mode for the number of expressed isoforms in genes with multiple isoforms was two (Figure 3.1). In our simulation, we randomly generated the number of isoforms we would simulate for each gene from a Poisson distribution with $\lambda = 2$. Since we were interested in only multiple isoform genes, we filtered this generated set of random numbers to include only values greater than one. Such a distribution had a mode of two isoforms, with a mean typically between 2 and 3 isoforms, which is similar to the distribution we saw in genes with multiple isoform in the TCGA LAML data set.

In order to make our simulations as similar to our real data sets as possible, we used `edgeR` (Robinson, McCarthy, and Smyth, 2010) to fit isoform counts from the TCGA lung adenocarcinoma (LUAD) data set (Cancer Genome Atlas Research Network, 2014) to a negative binomial distribution and estimated a mean and dispersion parameter for each isoform. Histograms of these estimates are seen in Figure 3.6.



(a) Case 1: Different gene expression

(b) Case 2: Different relative isoform usage

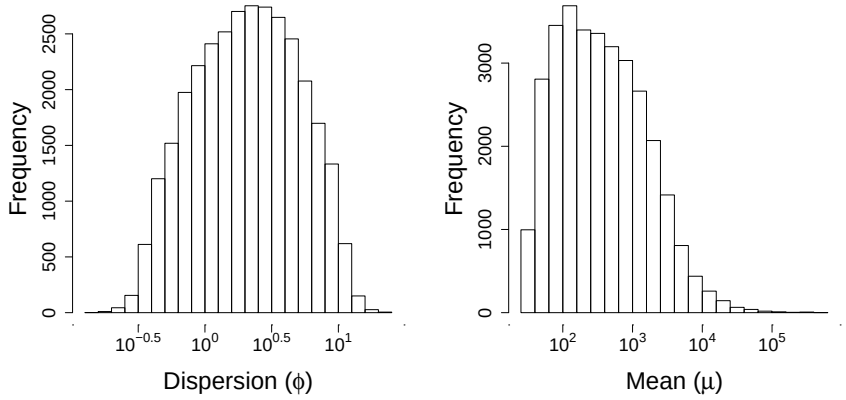
(c) Case 3: Genes and relative isoform usage varying within gene clusters

(d) Case 4: Genes and relative isoform varying across gene clusters

Figure 3.5: **Illustration of the different possible clusters simulated:** The cases shown in (3.5a) and (3.5b) depict when we allow for only gene clusters (g1-g3) or proportion clusters (p1-p9), respectively. The cases shown in (3.5c) and (3.5d) illustrate the two sets of clusters used when we combine the proportion clustering groups and the gene clustering groups in the same simulation. In both settings, the clusters showing differences in gene expression are the same as in (3.5a) (g1-g3), but the clusters showing differences in proportions differ. The case shown in (3.5c) illustrates the case where the nine proportion clusters (p1-p9) define subgroups of the three larger groups defined by gene expression differences (i.e. proportion groups are nested within gene groups). The case shown in (3.5d) illustrates the six clusters (p1-p6) used when the proportion groups can span the gene groups. Note that for this setting that while there are six groups showing differences in proportional isoform usage, the combination with differing gene expression levels mean there there are *nine* groups showing differences in isoform expression (i1-i9). Each of the small (nine) rectangles consists of 15 samples resulting in 135 samples.

In the description of our simulations, we noted that we consider the simulations to be the interaction of two different groups, the gene expression levels and the relative isoform frequencies. As such, we simulated these two factors separately to ensure control over the relationships in the groups. Briefly, we first simulated a set of relative isoform frequency vectors for each gene. Independent of those values, we then simulated a gene expression count for each gene.

Specifically, the first step in the simulation was generating a vector of relative isoform usage ratios for each gene. For each proportion group, we randomly sampled a set of parameters for each isoform from the list of mean and dispersion parameters estimated from the TCGA LUAD data. Using these parameters, we then simulated isoform counts for each sample within a proportion group by randomly generating values from a negative binomial distribution with the sampled mean and dispersion parameters. For each



(a) Dispersion Parameters (b) Means

Figure 3.6: **Histograms for estimates of parameters for simulation:** In order to simulate isoform count data similar to real data, we used estimates for the mean and dispersion parameter for isoforms from the TCGA lung adenocarcinoma dataset. These histograms show the distribution of those estimated parameters.

gene, we then used these isoform values to determine the proportional usage vector for that sample by dividing these isoform counts by the total sum of the isoforms within the gene.

In the second step of the simulation, we did not use these isoform totals to create the gene counts. Instead, the total gene count was simulated separately in the same manner. That is, for each isoform in a gene group, we again sampled parameter values from the distribution of values defined by the TCGA data. We then simulated isoform values for each sample from a negative binomial with those parameters. These isoform counts were then summed to get the gene estimates, per sample. The final isoform counts were derived by multiplying the gene counts by the final proportion vector.

3.2.3 Evaluating the Performance of Clustering

The simulated isoform counts, gene counts, and proportions were then clustered using the methodology described in Chapter 3.1.1 for values of K. Each simulation run consisted of 12 different starting data sets.

In order to quantify how well the clustering performed, we calculated the Jaccard similarity between the clusters we observed and the clusters we expected. For two cluster assignments A and B, where one is the expected clustering and one is the observed clustering, the equation for the Jaccard similarity is given as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN} \tag{3.6}$$

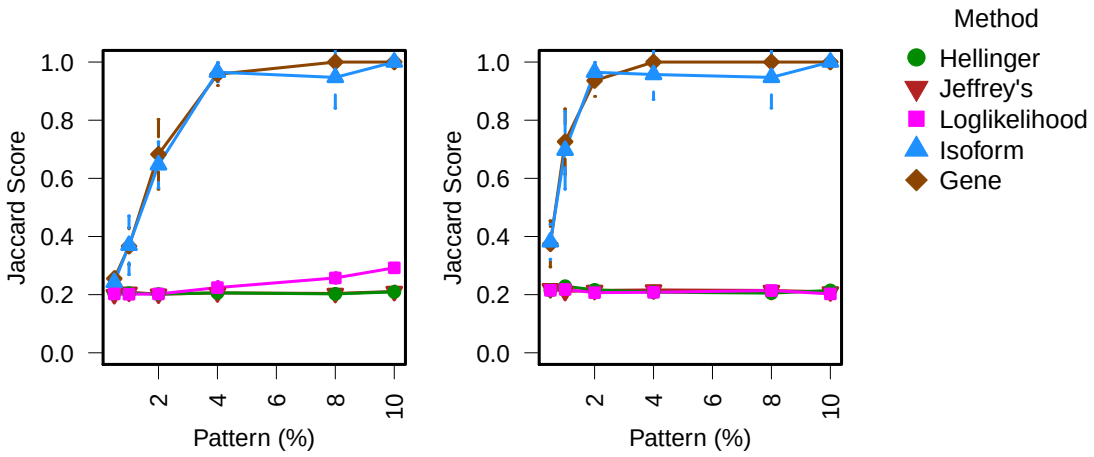
where TP, FP, and FN are defined as follows. The number of true positives, TP, is defined as the number of pairs that cluster together in the observed clusters as well as the true clustering. The number of false negatives, FN, is defined as the number of pairs that cluster together in the expected clustering but not the observed clustering. The number of false positives, FP, is defined as the number of pairs that cluster together in the observed clustering but not the expected clustering.

For each clustering, we calculated the Jaccard similarity that corresponded to the true signal(s) we expected the method to be able to detect. That is, we expect gene clustering to identify the gene groups and proportion clustering to identify the groups with differences in relative isoform frequency. We expect isoform clustering to identify groups with changes in either gene expression or relative isoform frequency. For example, in the case shown in Figure 3.5d, when the proportion clusterings span across the gene clusters, there are three gene clusters and six proportion clusters, while the intersection of these signals results in nine isoform clusters. Thus, we calculated how well the gene clustering at $K = 3$ performed at catching the three gene groups; how well the isoform clustering at $K = 3$ performed at catching the three gene groups as well as how isoform clustering at $K = 9$ performed at catching the nine isoform groups; and how well the proportion clustering at $K = 6$ performed at catching the six proportion groups. In the case illustrated in Figure 3.5c, the proportion clusters were nested within the gene and we were interested in how well the gene clustering with $K = 3$ caught the three gene groups; how well the isoform clustering with $K = 3$ performed at catching the three gene groups as well as how isoform clustering with $K = 9$ performed at catching the nine isoform groups; and how well the proportion clustering with $K = 9$ performed at catching the nine proportion groups.

3.3 Results of Simulations

We will consider each case discussed in Chapter 3.2.1. In Case 1, we consider only a change in gene expression (results seen in Figure 3.7). As expected, this signal is accurated detected by clustering on gene expression values. If we examine clustering isoform expressions directly, clustering on the isoform levels quite readily captures group differences when they are due to overall gene expression differences, even if the signal is only in around 2-4% of the genes. Additionally, proportion based methods do not capture this signal, as expected.

In Case 2, we consider genes showing only a change in relative isoform usage. The results of the simulations show that proportion based clustering reliably finds the clusters that vary by relative isoform usage even when the pattern represents a low percentage of genes in the data (around 2% of genes, Figure 3.8). In contrast, clustering on isoform expression does not perform as accurately at this fraction of genes with the signal. In this scenario where there is no competing gene signal and all clustering is due to relative isoform usage, the percentage of the genes with the signal must be much higher (around



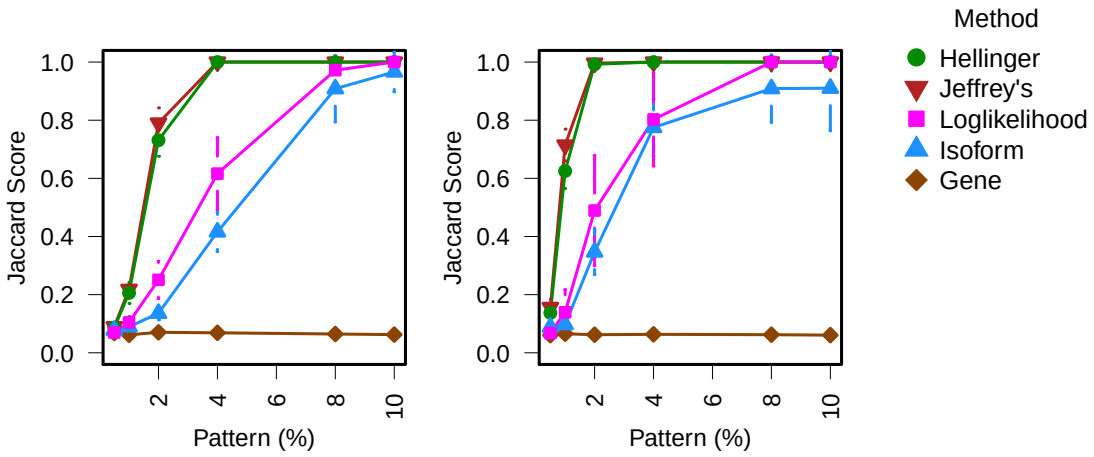
(a) K-medoids Clustering (b) Hierarchical Clustering

Figure 3.7: **Gene expression varies while relative isoform usage remains constant (Case 1):** The results of clustering when gene expression values are different across the three gene groups (see Figure 3.5), while the relative isoform usage within these groups remains constant. The x-axis gives the percent of genes that show this clustering pattern, while the remaining genes are held constant across all samples. Gene and isoform clustering differentiate the expected three groups and perform quite similarly. As expected, proportion clustering does not distinguish the three gene groups.

8% of genes) than that seen with the proportion based methods in order for this method to find the correct groups. As expected, gene clustering is unable to find these groups.

In both Case 3 and Case 4, the gene and relative isoform usage groups both vary but differ in the relationship between gene levels and relative isoform usage. Specifically, in Case 3, the relative isoform usage groups are subsets of the differential gene expression groups, while in Case 4, some of the relative isoform usage group spans the differential gene expression group. The results of considering Case 3 and Case 4 by themselves are seen in Figure 3.9.

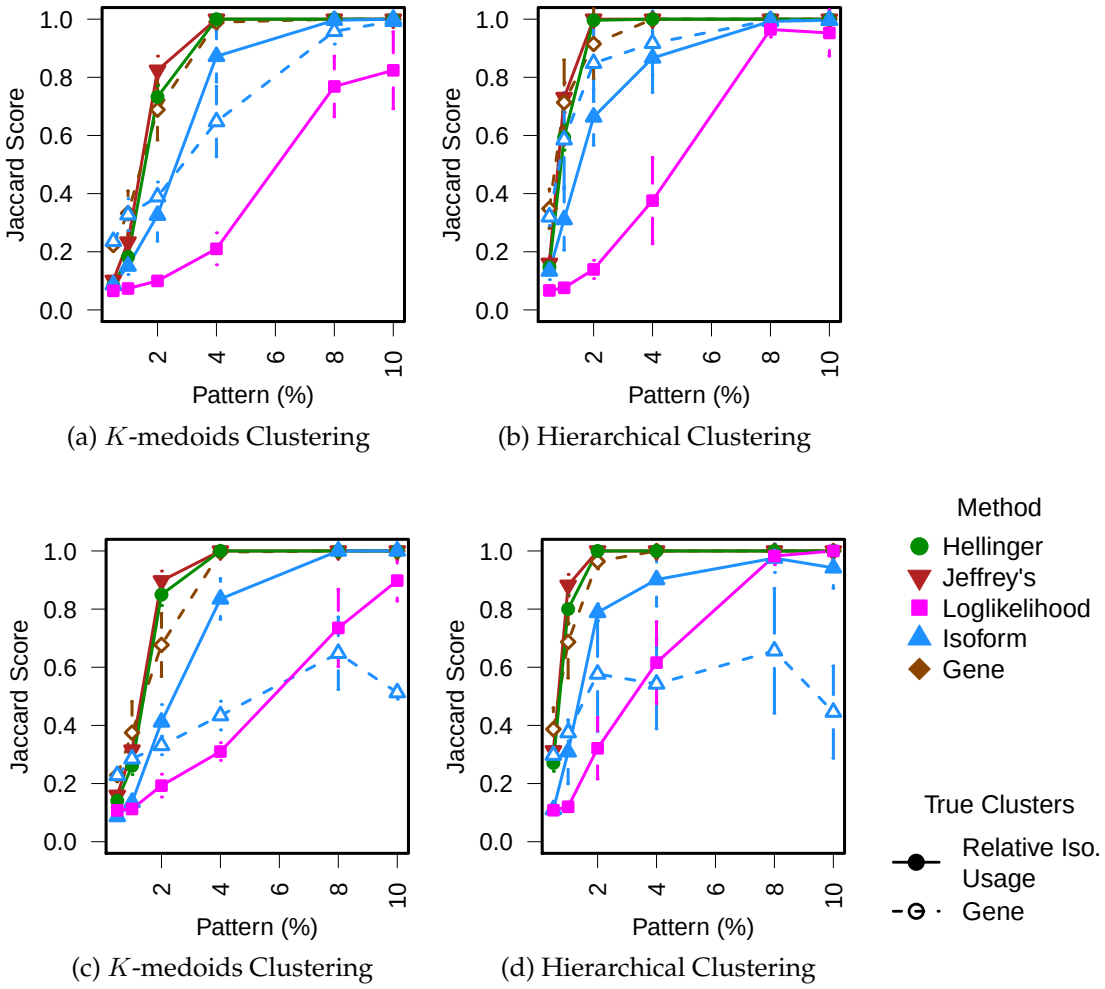
In this more complicated setting, the proportion based methods still identify the correct relative isoform usage clusters even at very low frequency. The isoform clustering generally finds the differential gene clusters and shows some improvement at catching the relative isoform usage signal as the gene expression levels provide information about this clustering as well. However, isoform clustering does not perform as well as clustering on the proportional isoform usage. Clustering on the isoform expression only starts to have high concordence with the true proportional usage clusters when about 7-10% of the genes show that pattern – a much higher required percentage than that of the proportion clustering, which finds the same pattern reliably even when only 2% of the genes show the pattern.



(a) K-medoids Clustering (b) Hierarchical Clustering

Figure 3.8: **Relative isoform usage varies, gene expression remains constant (Case 2):** The results of clustering when the relative isoform usage is different across the nine clustering groups, while the gene expression values remains the same for all genes (Figure 3.5b). The x-axis gives the percent of genes that show the proportion clustering pattern, with isoform usage constant across all samples in the remaining genes. Proportion clustering readily identifies the nine groups, while isoform clustering does so only once a large percentage of the genes show the pattern. As expected, gene clustering does not distinguish the nine groups.

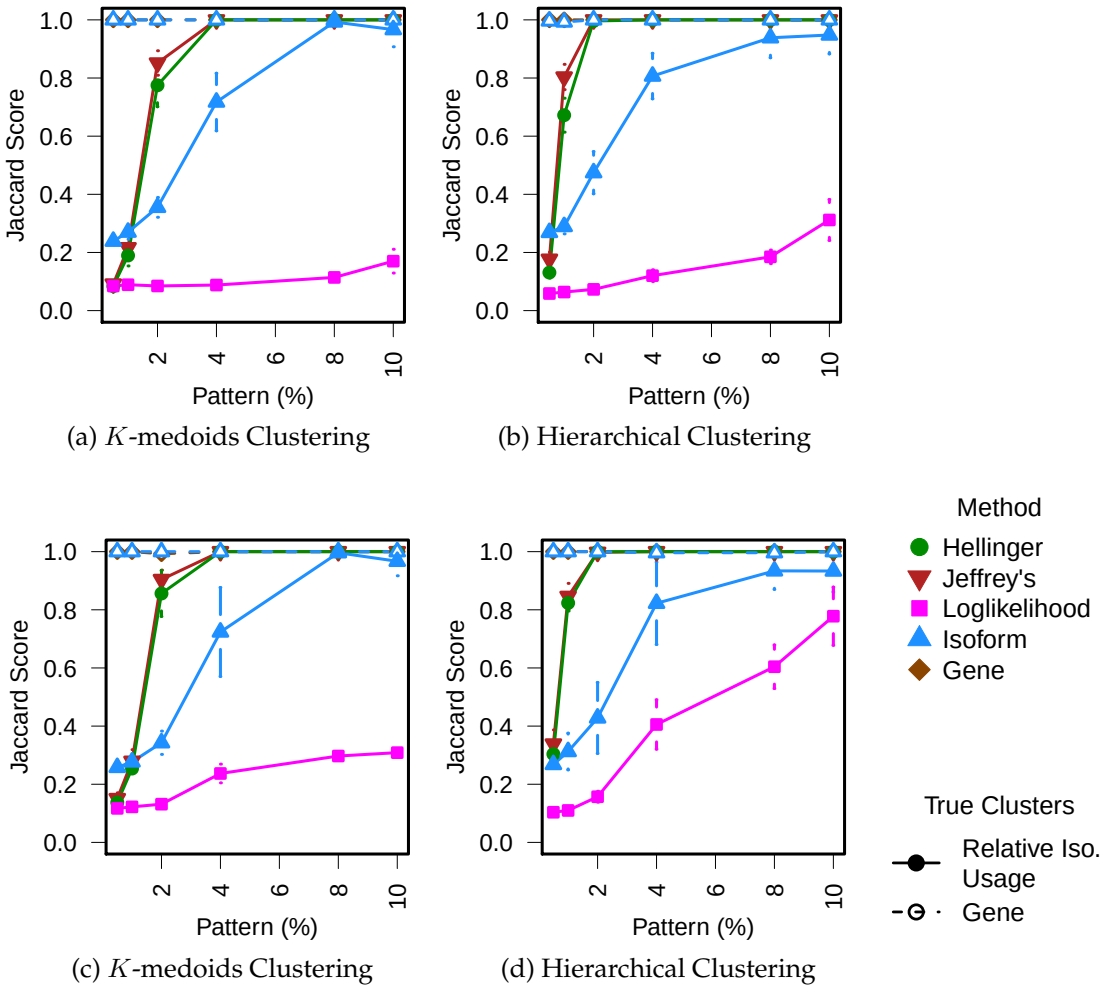
Moreover, the proportion clustering finds the relative isoform usage pattern even in the presence of complicated backgrounds of other clustering signals based on gene expression differences (Figure 3.10). In the previous results, the gene expression levels of genes not in the groups were simply noise, without any type of clustering signal. The same is also true for the changes in relative isoform frequency. That is, for genes not in the simulated groups, the level of isoform usage remained constant across all genes. However, we additionally wanted to simulate a case where there was an additional significant clustering signal. In this case, we chose 25% of the background genes to also have gene expression differences while maintaining the small number of genes having both gene and relative isoform frequency shifts. In real data sets, we expect that differences in alternative splicing will affect a comparatively small numbers of genes and differences in overall gene expression will dominate. These result indicate that proportion clustering has the potential to be more sensitive to finding clustering based on this type of alternative splicing, and particularly when there are a mix of gene and alternative splicing signals as would be expected in true biological data.



(a) K-medoids Clustering (b) Hierarchical Clustering

(c) K-medoids Clustering (d) Hierarchical Clustering

Figure 3.9: **Gene expression and relative isoform usage both vary within gene (Case 3 and Case 4):** The x-axis gives the percent of genes that show the proportion clustering pattern. The relative isoform usage remains constant across all samples in the remaining genes. For (3.9a) and (3.9b), the true proportion clusters are nested within the gene groups. In (3.9c) and (3.9d), the proportion clusters span the gene groups. When the proportion groups span the gene groups, the number of correct groups to find differs between proportion clustering (6) and isoform clustering (9).



(a) K-medoids Clustering (b) Hierarchical Clustering

(c) K-medoids Clustering (d) Hierarchical Clustering

Figure 3.10: **Variation across genes (combining Case 1 with either Case 3 or Case 4):** Here we repesent a likely biological scenario where most of the clustering signal is due to genes showing only gene expression differences (Case 1), with a smaller proportion of genes also show proportion and gene expression differences (Case 3). The percent of genes that show both gene and proportion differences is allowed to vary (shown by the x-axis), while a fixed 25% of the genes show only gene expression patterns (Case 1). The remainder of the 5,000 genes are held constant across all samples in both gene expression and relative isoform usage. For (3.10a) and (3.10b), the true proportion clusters are nested within the gene groups (Figure 3.5c), while in (3.10c) and (3.10d), the proportion clusters span that of the gene groups (Figure 3.5d). When the proportion groups span the gene groups, the number of correct groups to find differs between proportion clustering (6) and isoform clustering (9).

3.4 Sparse Clustering

The features that are responsible for the true clusters found between individuals are likely only a small subset of the features being examined. Sparse clustering may improve the ability to find the true clusters in the data as some amount of non-informative data will be removed. Additionally, sparse clustering also potentially leads to improved interpretability of clustering results since the clustering methodology would also be informative as to which features are driving the clustering.

3.4.1 Description of Sparse Clustering

Much of the initial framework for our clustering method came from work by Witten and Tibshirani (2010), which is summarized here. Many clustering methods can be viewed as an optimization problem which can be written in the form

$$\underset{\Theta \in D}{\text{maximize}} \left\{ \sum_{j=1}^{p} f_j(\mathbf{X}_j, \mathbf{\Theta}) \right\}$$

In this notation, $f_j(\mathbf{X}_j, \mathbf{\Theta})$ is some function which only involves the jth feature of the data and $\mathbf{\Theta}$ are parameters in set D. As an example, we can consider how K-means clustering fits into this framework. The measure of distance used in K-means clustering is typically the Euclidean distance, and thus, we can define $d_{i,i',j} = (X_{ij} - X_{i'j})^2$. The objective function in K-means clustering is to minimize the within cluster sum of squares (WCSS).

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} \sum_{j} d_{i,i',j} \right\}$$

However, minimizing the WCSS would be the same as maximizing the between cluster sum of squares (BCSS). Therefore, we want to maximize

$$\underset{C_1,\ldots,C_K}{\text{maximize}} \left\{ \sum_{j=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j} \right) \right\}$$

We can see in this framework, for the case of K-means clustering, f_j is the the between cluster sum of squares for feature j, and $\mathbf{\Theta}$ is a partition of the observations into K disjoint sets.

Further, in order to achieve sparse clustering, the optimization problem is modified to

$$\underset{\Theta \in D}{\text{maximize}} \left\{ \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j, \mathbf{\Theta}) \right\} \text{ subject to } ||\mathbf{w}||^2 \leq 1, ||\mathbf{w}||_1 \leq s, w_j \geq 0 \; \forall j.$$

In this notation, w_j is the weight for feature j. The value of w_j can be interpreted as the contribution of feature j to the resulting clustering. That is, a large value of w_j indicates a feature that is driving the clustering, while $w_j = 0$ means that feature j is not involved in the clustering. Additionally, a tuning parameter s limits the summation of the weights. In the **sparcl** implementation of this, optimization is performed by soft-thresholding (Witten and Tibshirani, 2010).

If we continue the example using K-means clustering, in order to perform sparse K-means clustering, we are attempting to solve the problem

$$\underset{C_1,\ldots,C_K,\mathbf{w}}{\text{maximize}} \left\{ \sum_{j=1}^{p} w_j \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{i,i',j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{i,i',j} \right) \right\}$$

$$\text{subject to } ||\mathbf{w}||^2 \leq 1, ||\mathbf{w}||_1 \leq s, w_j \geq 0 \; \forall j.$$

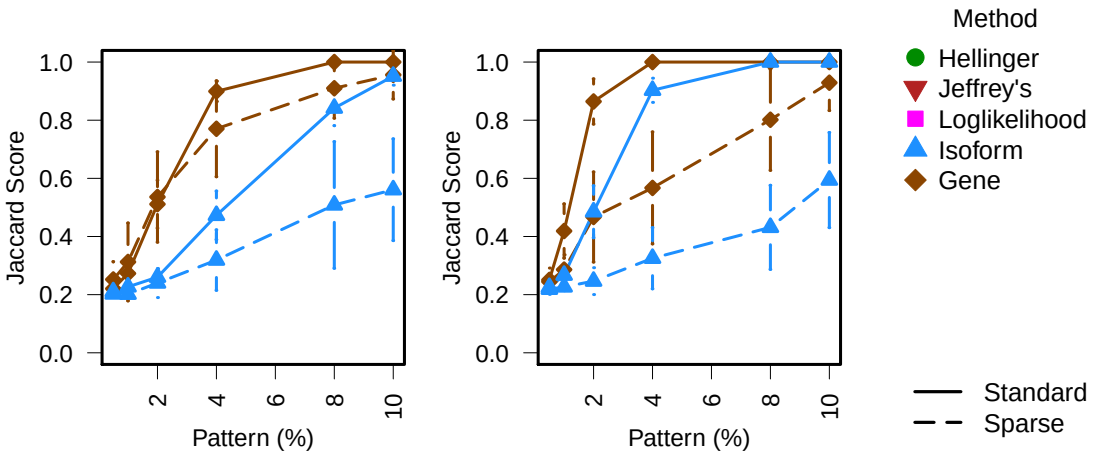
On both simulated and genomic data sets, Witten and Tibshirani (2010) described this sparse clustering objective outperforming both standard clustering techniques as well as other clustering methods using feature selection. The metric used to measure accuracy was the clustering error rate (CER) which was discussed in Chapter 1. The simulations included a small simulation with n=5 and p=25 as well as a larger simulation of n=50 and p=500. Additionally, a breast cancer data set with four subgroups was also analyzed (Perou et al., 2000). This dataset had expression information for 1753 genes in 62 patients. Again, sprase clustering method was more accurate than standard clustering for grouping the data.

3.4.2 Evaluating the Performance of Sparse Clustering

In order to perform sparse clustering, we began with the code from the R package **sparcl** described by Witten and Tibshirani (2010). In our implementation, we modified the code in this package to function as described in Chapter 3.1.1. That is, isoform counts were transformed into gene counts and proportions. For each of the features, individual distance matrices were determined. The code was modified as necessary to use distance matrices rather than that data matrix X itself. After finding weights for each of the distance matrices for the features, the overall distance matrix, now sparse in its features, was used to perform the final clustering.

The simulations were run as described in Chapter 3.2.2. However, due to memory and time constraints, these simulations were performed on 1,000 rather than 5,000 genes. We note that the standard clustering results are mostly consistent from results we described earlier in Section 3.3. Therefore, this section will mostly focus on the results specific to the comparison in accuracy of sparse and standard clustering.

We began by considering the simplest cases, either the data had only gene expression differences or the data had relative isoform expression differences, but not both. Figure 3.11 shows the case where the differences in the known groups are due only to differences in gene expression. This simulation is quite similar to that seen in the original work



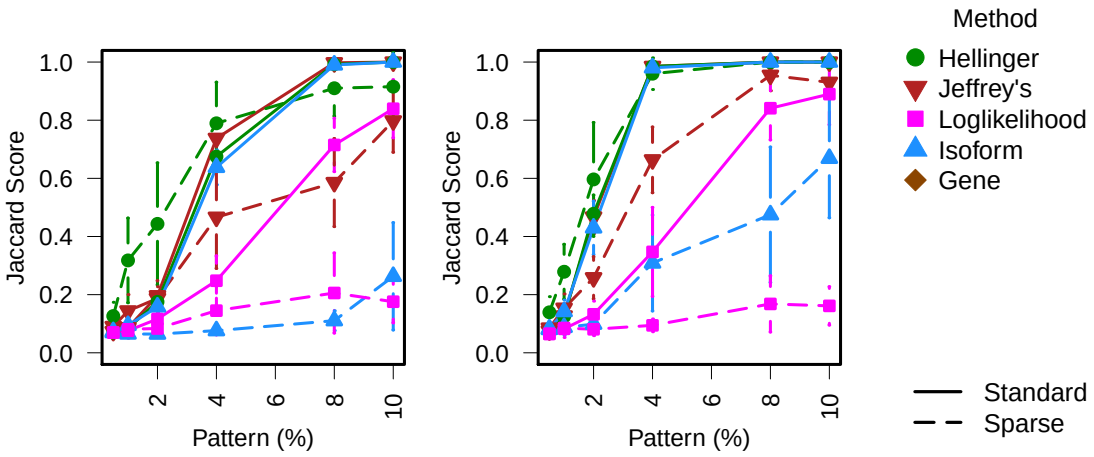
(a) K-medoids Clustering (b) Hierarchical Clustering

Figure 3.11: **Variation across gene groups (comparing Case 1 in standard and sparse clustering):** These simulations show the results of both sparse clustering and standard clustering when the gene expressions values vary across the the gene expression groups (see Figure 3.5), while the relative usage of the isoforms in these genes remains constant. This simulation was similar to the simulation in Figure 3.7, except we simulated 1,000 genes in this instance. The x-axis gives the percent of genes that show this clustering pattern, with the remaining genes have constant expression across all samples. As we saw in Figure 3.7, only gene and isoform clustering differentiate the expected three groups, so only those clustering results are plotted here.

described by Witten and Tibshirani (2010), which performed clustering on gene counts. However, somewhat surprisingly, we saw that the average Jaccard score obtained from sparse clustering did not improve compared to the clustering results from standard clustering in either K-medoid or hierarchical clustering.

Figure 3.12 shows the case where the only differences between clusters are due to relative isoform usage with no underlying difference in gene expression. When we clustered using isoform expression, relative isoform usage using Jeffrey's divergence, or the Log-Likelihood based distance, we saw that standard clustering results were more accurate results than sparse clustering in both K-medoid and hierarchical clustering. However, this is not the case in clustering on relative isoform usage using Hellinger's distance. In this instance, we saw more accurate results using sparse clustering than standard clustering in both K-medoid and hierarchical clustering.

Figures 3.13a and 3.13b show clustering when clusters contain both a differential gene expression signal as well as an differential isoform usage signal. If we look at gene and isoform expression clustering, for the most part, we see that standard clustering is actually outperforming sparse clustering in these simulations. In some cases, when the



(a) K-medoids Clustering (b) Hierarchical Clustering

Figure 3.12: **Variation across relative isoforms (comparing Case 2 in standard and sparse clustering:** The results of clustering when the relative isoform usage varies across the nine clustering groups, but gene expression values remain constant for all genes (Figure 3.5b). This simulation was similar to the simulation of Figure 3.8, except we simulated 1,000 genes in this instance. The x-axis gives the percent of genes that show the proportion clustering pattern, with the remainder of genes displaying constant relative isoform usage across all samples. As we saw in Figure 3.8, gene clustering will not differentiate the expected nine groups and is not shown here.

pattern is only present in a very small number of genes, we do see instances where the Jaccard score is higher in sparse clustering than standard clustering. However, due to the variability in both the standard and sparse clustering, we are unable to say that sparse clustering is more accurate at very low percentages. At higher percentage of pattern in the genes, we see clear separation of the average Jaccard score in sparse and standard clustering, and standard clustering is typically more accurate in both hierarchical clustering and K-medoid clustering.

The results for proportion based clustering in seen in Figures 3.13a and 3.13b. We see with the Log-Likelihood distance and Jeffrey's divergence that standard clustering is more accurate than sparse clustering. However, we again note that clustering with the Hellinger distance performs better, particularly in K-medoid clusterings, when used in conjunction with sparse clustering than standard clustering.

In general, we did not see the same strong results from sparse clustering that were seen in the original paper describing the methodology (Witten and Tibshirani, 2010). Not only were the clusterings less accurate, but we also noticed a lot of variability in our results. In particular, we noticed the number of features with non-zero weights could vary considerably in simulations using the same parameters. For example, we noticed

that different runs of simulation (that is, the same number of genes with a pattern) under the same parameters would vary from having all non-zero features in one run to a very few non-zero features in another run. The number of non-zero features is dependent on a tuning parameter which is selected using the gap statistic, which Witten and Tibshirani (2010) mentions does not always perform consistently well. Additionally, in comparing our simulations to those presented in Witten and Tibshirani (2010), we note that their simulations were performed on a smaller number of total features and the percentage of features with the cluster pattern was larger than in our simulations.

3.5 Relationship of Distance to Kernel Methods

Our clustering framework is described as calculating an overall distance matrix from the combination of distance matrices over many features. Because of the relationship between distances and kernels, we can see a relationship between our clustering method and kernel methods. Kernel methods are useful when data in the input space does not have a linear relationship but does have a linear relationship if mapped to a higher-dimensional space. This higher dimensional space is termed a feature space, and in this feature space, it is simpler to apply linear algorithms to the mapped data.

More explicitly, let the data be written as $\mathbf{x}_i, \mathbf{x}_j \in X$. Additionally, we have a mapping between the input space and the feature space $\phi : X \to \mathbb{R}^N$. We may now define the kernel, K as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle.$$

This measure, the kernel, is often interpreted as a similarity measure.

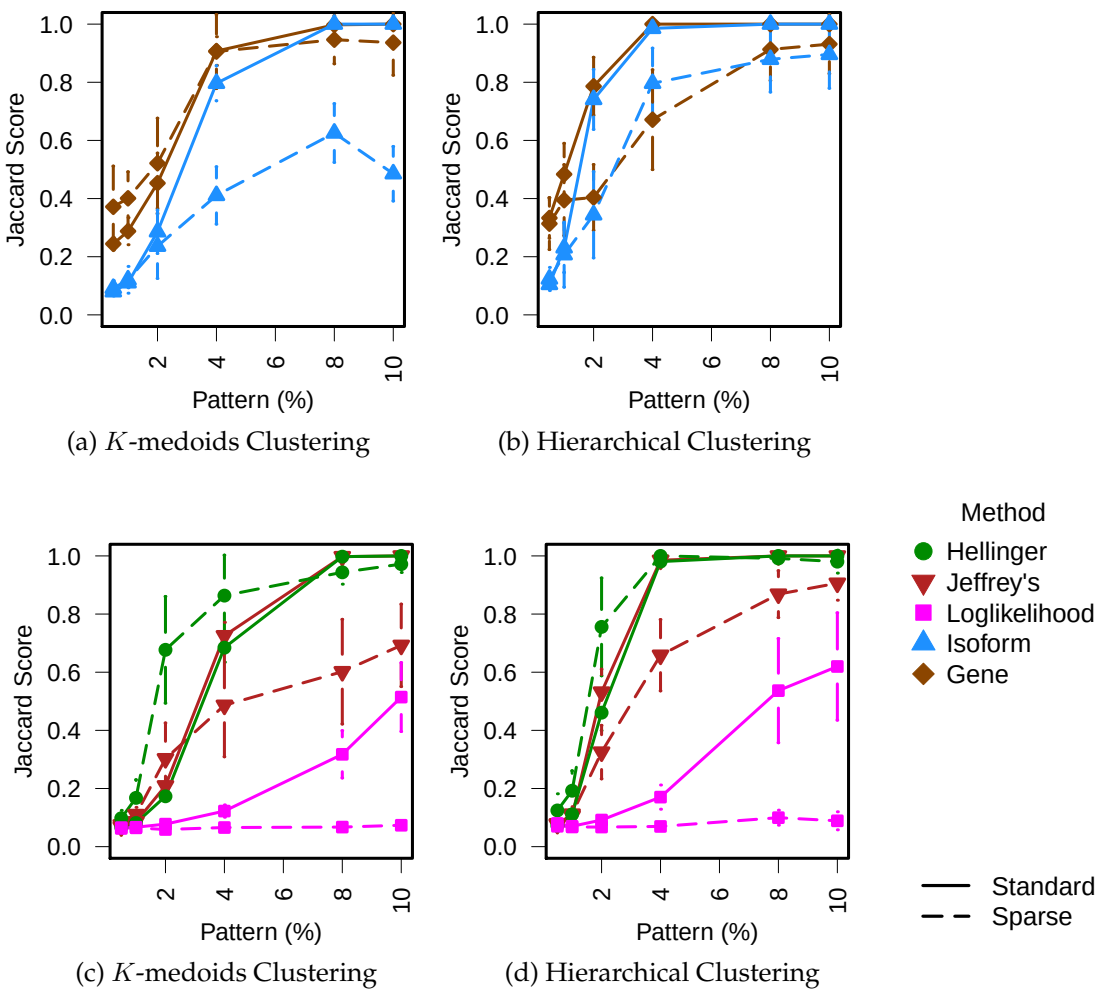
Because X is a reproducing kernel Hilbert space, it has a natural notion of distance, namely $||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2$. A useful feature of kernel methods is that it is possible to compute distances in the feature space without knowing the mapping ϕ. This can be done using the distance kernel trick:

$$||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 = K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j)$$

We can see the distance in the feature space can be calculated from the input vectors. It is possible to use kernel methods on any algorithm in which input vectors appear only as dot products with other input vectors. For simplication of notatation, we can define the Gram matrix K as the matrix where the entry k_{ij} is the scalar product $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. Therefore, the Euclidean distance may also be written as

$$||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 = k_{ii} + k_{jj} - 2k_{ij}$$

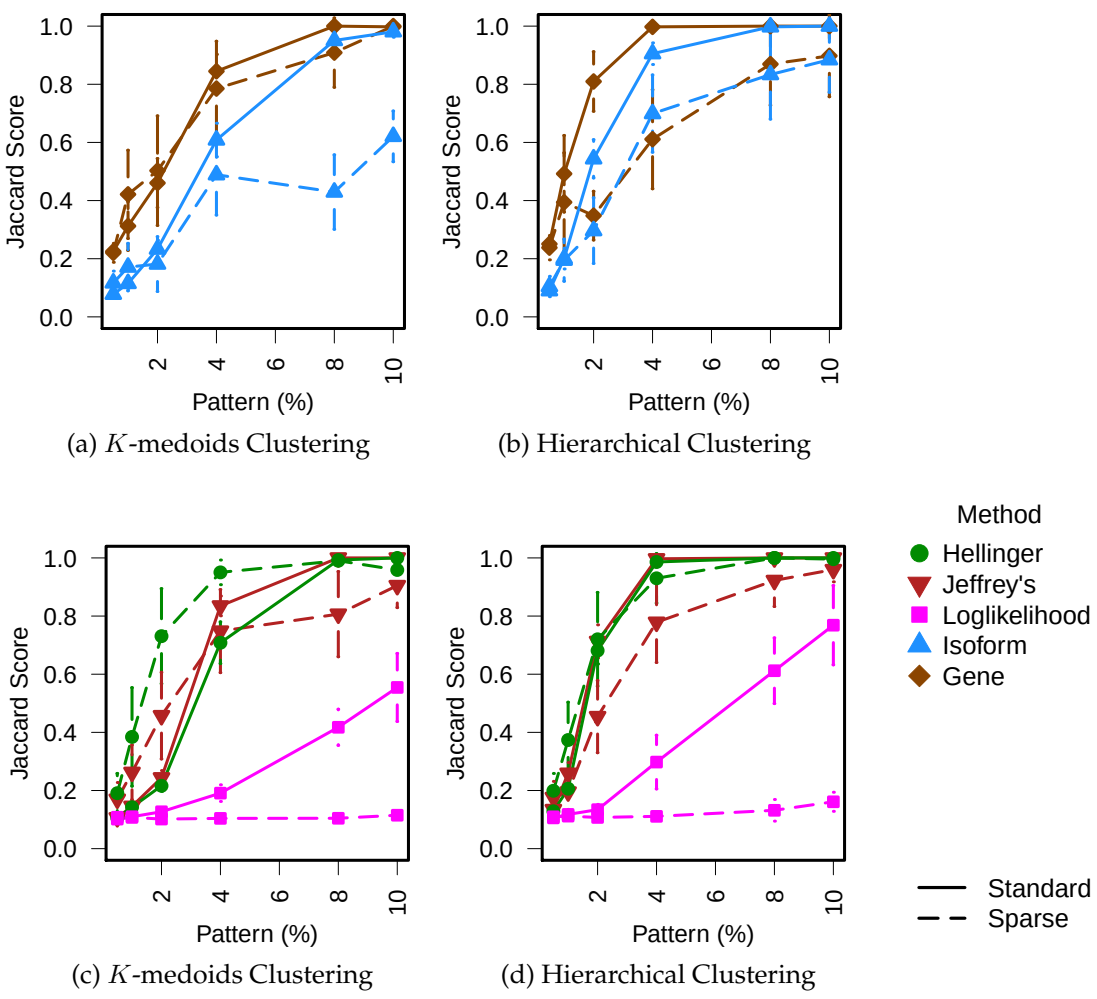
The support vector machine (SVM) used in classification problems is arguably the most well known kernel-based algorithm. The success of SVM led to the use of kernel methods in other learning algorithms, including unsupervised learning algorithms such



(a) K-medoids Clustering (b) Hierarchical Clustering

(c) K-medoids Clustering (d) Hierarchical Clustering

Figure 3.13: **Variation across genes (combining Case 1 with Case 3 using sparse clustering:** Here we repesent a likely biological scenario where most of the clustering signal is due to genes showing only gene expression differences (Case 1) while a smaller proportion of genes also show proportion and gene expression differences (Case 3). This simulation set-up is similar to the simulation in Figure 3.10, though these simulations used 1,000 rather than 5,000 simulated genes. The percent of genes that show both gene and proportion differences is allowed to vary (shown by the x-axis), while a fixed 25% of the genes show only gene expression differences with constant relative isoform usage (Case 1). The remainder of the 1,000 genes are constant across all samples in both gene expression and relative isoform usage. In these cases, the true proportion clusters are nested within the gene groups.



(a) K-medoids Clustering (b) Hierarchical Clustering

(c) K-medoids Clustering (d) Hierarchical Clustering

Figure 3.14: **Variation across genes (combining Case 1 with Case 4 using sparse clustering:** Here we repesent a likely biological scenario where most of the clustering signal is due to genes showing only gene expression differences (Case 1), while a smaller proportion of genes also show proportion and gene expression differences (Case 3). This simulation set-up is similar to the simulation in Figure 3.10, though these simulations used 1,000 rather than 5,000 simulated genes. The percent of genes that show both gene and proportion differences is allowed to vary (shown by the x-axis), while a fixed 25% of the genes show only gene expression patterns with constant relative isoform usage (Case 1). The remainder of the 1,000 genes are constant across all samples in both gene expression and relative isoform usage. In these cases, some of the true proportion clusters span the gene groups.

as prinicipal component analysis, canonical correlation analysis, and clustering. For example, clustering could be performed by mapping the data to a feature space and calculating the centroids in this mapped feature space. The distance between each element and the centroid in the feature space could then be found using the distance kernel trick (Filippone et al., 2008).

Additionally, kernels can be combined together to form a more complex kernel. Specifically relevant to the description of our clustering method, we know that the kernels display the following properties:

1. $K(\mathbf{x}_i, \mathbf{x}_j) = K_1(\mathbf{x}_i, \mathbf{x}_j) + K_2(\mathbf{x}_i, \mathbf{x}_k)$
2. $K(\mathbf{x}_i, \mathbf{x}_j) = \alpha K_1(\mathbf{x}_i, \mathbf{x}_j)$, where $\alpha \in \mathbb{R}^+$

Therefore, we may calculate a new kernel based on the weighted average of other kernel. This is similar to our method of finding an overall distance matrix from the weighted average of the distance matrix of each feature. Our approach can be seen as defining a separate kernel for each feature and then combining the kernels via (weighted) averaging. Methods for combining multiple kernels have been proposed and how to choose functionals that combine multiple kernels is termed the multiple kernel learning problem (see Gönen and Alpaydın, 2011, for a review), and weighted combinations like we describe are common. This idea of combining kernels has already been used in genomic studies. For example, kernel methods have been used in gene expression studies previously in order to combine gene expression profiles with protein interaction networks (Lavi, Dror, and Shamir, 2012). Since many of these methods have computational difficulties for the large numbers of features we have here, Zeng and Cheung (2011) propose creating a sparse multiple kernel for each feature for unsupervised clustering, with the similar goal as described in Witten and Tibshirani (2010) of finding sparse weights for the linear combinations of the kernels for each feature.

Chapter 4

Real Data: Identifying Batch Effects

4.1 Finding a Gold Standard

The simulations presented in Chapter 3 allowed us to test settings composed of signals for both changes in gene and differential isoform usage. We saw in those simulations that we were able to differentiate each signal using the appropriate clustering technique. For the specific type of data we are interested in, that is, observations showing shared differential isoform usage, it is difficult to compare clustering methods on real data as we do not often know true groups. Ideally, we would utilize real data sets for testing our method that would return different results for gene expression based clustering and differential isoform usage based clustering. Potentially, such a case could be isoforms regulated by the same promoter or splicing factors. However, we were unaware of a well established model to use for this specific purpose.

We hypothesize based on what we know about alternative splicing that we should find differential isoform usage between different tissue types as well as between tumor and normal samples. However, these type of data sets are not ideal for testing our method as these changes in differential isoform usage are also typically manifested as changes in gene expression. As a result, clustering on gene expression would return the same result as clustering on relative isoform usage.

4.2 Using Batch Effect As Gold Standard

TCGA is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) which has been responsible for generating several types of -omics data for 33 cancer types. Typically, for each tumor type, several different types of data were generated, including exome sequence, mRNA sequencing, microRNA sequence, DNA copy number, and DNA methylation.

Among the findings of the exome sequencing of several of these cancer types were instances of recurrent splicing mutations. Among the cancer types these somatic muta-

tions were found in were LAML (Cancer Genome Atlas Research Network, 2013), LUAD (Cancer Genome Atlas Research Network, 2014), and breast cancer (Cancer Genome Atlas Research Network, 2012). While some of these recurrent mutations were previously known and studied, an advantage of the TCGA project is the ability to further utilize the corresponding mRNA sequencing data to explore the splicing alterations that result due to mutations. Additionally, the larger sample sizes recruited by the TCGA project potentially increased the variability of phenotypes seen as well as increased the ability to find rare subpopulations (Song, Merajver, and J. Z. Li, 2015).

Due to the large scale, multi-center nature of projects like TCGA, research projects may include data that has been collected or processed by many different centers. For example, tumor samples are collected at various Tissue Source Sites (TSS). These are transported to Biospecimen Core Resources (BCR) laboratories to be processed and stored. The specimens are grouped into batches and sent on to sequencing centers for sequencing. In some cases, some of the tumor types have had hundreds of samples collected. Simply due to volume of samples, it would be necessary for the samples to be collected at various TSS sites, processed at a several BCR laboratories, and sequenced on many plates. However, experimental variation frequently occurs across batches of experiments that have been performed by different centers, by different personnel, or even on different days (Leek et al., 2010). Part of the pipeline for all TCGA analyses was to examine the data for the presence of possible batch effects. Among the variables investigators examined as a potential source of batch effects were the TSS, the batch ID, the shipment date to the sequencing center, and the plate ID.

One of the data sets we chose to study was the mRNA-Seq data of LAML tumors. Originally, we selected this data set due to the presence of somatic mutations in a gene, *U2AF1*, that encodes known splicing factor (Cancer Genome Atlas Research Network, 2013). In their analysis of the data, the conclusion of the TCGA working group was that no serious batch effects were present in the LAML mRNA-Seq data. They concluded that technical batch effects in the data set were reasonably small and unlikely to influence analyses in any important way and as such, did not require any type of correction (Cancer Genome Atlas Research Network, 2013). However, in our own analysis of this data, we discovered a batch effect which strongly impacted our clustering analysis. As we will explain further, we showed that this batch effect manifested as different relative levels of isoform abundances within genes. This gives us a gold-standard to which we can compare the accuracy of our clustering methods.

4.2.1 Identification of 5′ to 3′ bias

Early analysis of the LAML data set suggested that despite the LAML working group's conclusion, there was indeed a batch effect present in this data sets. One potential source of the batch effect was suggested to be a 5′/3′ bias. The 5′/3′ bias is a well documented bias of mRNA-Seq data (Sigurgeirsson, Emanuelsson, and Lundeberg, 2014; Wu, X. Wang, and Zhang, 2011; Mortazavi et al., 2008) where the 3′ end of a transcript is more likely to

be captured and sequenced than the 5′ end of a transcript. Various events in the library preparation can cause such an effect, including partial degradation of mRNA molecules, (Wu, X. Wang, and Zhang, 2011), so it is plausible that small variation in library preparation between the plates could cause this difference. With respect to relative isoform usage, isoforms often differ in their starting and ending exons, and therefore different relative coverage of the beginning or end of a gene due to technical artifacts can mean that isoforms will get assigned different relative expression levels. Indeed, many methods for estimating isoform expression have been proposed to model this bias (Roberts et al., 2011; W. Li and T. Jiang, 2012; Wu, X. Wang, and Zhang, 2011).

In order to determine whether a 5′ or 3′ bias is present, we implemented the Python module `geneBody_coverage.py` found as part of RSeQC (`http://rseqc.sourceforge.net/`) (L. Wang, S. Wang, and W. Li, 2012). This module calculates the read coverage across genes to determine if read coverage is uniform or contains bias in coverage across the genes. RSeQC divides the gene into equally spaced bins and calculates the number of sequences falling in the bin, relative to the overall number sequences assigned to the gene region. For more details see L. Wang, S. Wang, and W. Li (2012).

This module requires raw BAM files as input, which we downloaded from the Cancer Genomics Hub (`https://cghub.ucsc.edu/index.html`). The raw BAM files (unlike the count summaries) are held under controlled access and only available to those who have applied for and received a Data Access Request (DAR). Additionally, a BED file describing a set of housekeeping genes is also required as input to RSeQC and is available on the RSeQC site: `http://sourceforge.net/projects/rseqc/files/BED/Human_Homo_sapiens/hg19.HouseKeepingGenes.bed`.

We limited the housekeeping genes to those which showed expression in only one isoform in our dataset (that is, were single isoform genes) so as to avoid any issue of differential coverage due to alternative splicing. We separated the bed file by the size of the gene and looked separately at genes with less than 1,000 base pairs (318 genes), 1,000-1,500 base pairs (460 genes), 1,500-2,000 base pairs (532 genes), 2,000-2,500 base pairs (493 genes), and 2,500-3,000 base pairs (474 genes).

When we examined the plate effect closely, we see that plate 734 appears to have a different level of 5′/3′ bias. In Figure 4.1, we show a plot of the relative proportion of the mRNA-Seq sequences that came from the beginning of the transcript (5′ end) versus the end of the transcript (3′ end), as calculated by RSeQC (L. Wang, S. Wang, and W. Li, 2012). It is clear from this plot that plate 734 has greater relative coverage of the beginning of the gene compared to the other plates. Figure 4.1 shows short genes (0-1000 base pairs) but the effect can be seen in a range of gene lengths.

4.2.2 Implementation of Clustering

The identification of this 5′/3′ bias in the TCGA LAML data set suggested that we had found a gold standard data set which showed differential isoform usage between group. In order to test our differential isoform usage clustering method, we performed gene,

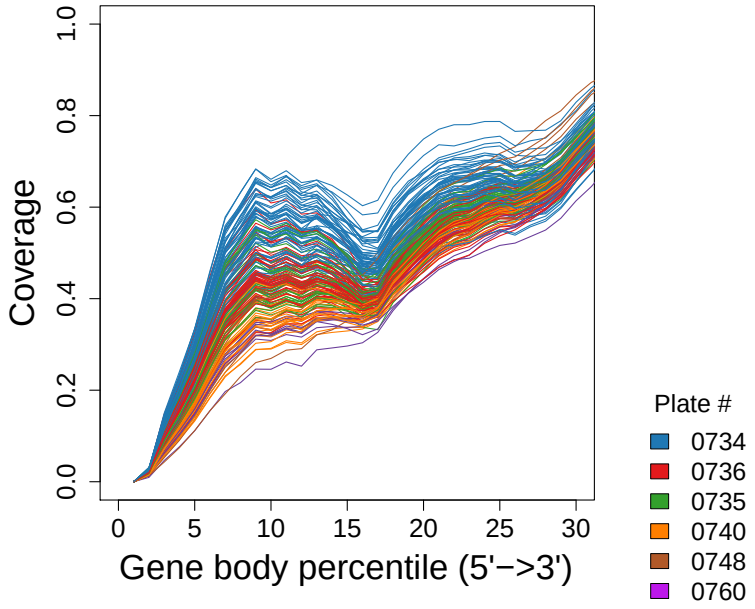


Figure 4.1: **5′ to 3′ bias differs by plate.** Here we show the results of RSeQC (L. Wang, S. Wang, and W. Li, 2012) calculation of the average coverage of the mRNA-Seq data; shown here are the results of the calculation for 318 housekeeping genes that have only a single isoform and have total length in the range of 0-1000 base pairs. This calculation divides the gene into equally spaced bins and calculates the number of sequences falling in the bin, relative to the overall number sequences assigned to the gene region. The x-axis shows the percentile of the gene body that the bin falls in (referenced from the beginning, or 5′ end, of the gene). This plot shows a closeup of the results at the 5′ start of the gene.

isoform, and proportion clustering on counts downloaded from the TCGA data portal (download via `https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm` on 10/14/14). Specifically, we downloaded the RNASeqV2 level 3 raw isoform counts from the portal. The TCGA data pipeline used to generate these counts included alignment to the reference genome using Mapsplice (K. Wang et al., 2010) and quantitation of transcripts using RSEM (B. Li and Dewey, 2011). RSEM returns two kinds of estimates, which include an estimate of the number of fragments that are derived from a given isoform or gene as well as an estimate of relative expression called transcripts per million (TPM). For our purposes, we used the estimates of the number fragments (counts) derived per isoform. We used the annotation file provided by TCGA to associate isoforms with genes, which can be found at `https://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf`.

The TCGA LAML dataset contained expression data for 73,599 isoforms in 173 individuals. To cluster the AML data, we normalized the AML samples by TMM normalization (Robinson, McCarthy, and Smyth, 2010) and performed an initial filtering of the data

to remove extremely lowly expressed isoforms. The low expression filter was definied as those isoforms that had a median value less than 25 counts across all samples. Using the isoform counts as our initial input, we created the three different types of features: gene counts, isoform counts, and isoform proportions. The number of isoforms and genes provided is still quite large – 28,014 expressed isoforms and 12,218 expressed genes.

We then performed variance filtering in order to reduce the size of the datasets. The gene count and isoform count datasets were filtered by selecting the 5,000 most variable genes or isoforms. We then calculated the distance matrix for each of these features. In the case of the isoform proportions, we first calculated the distance matrix for all features (i.e. genes). For the AML data, we focused on Jeffrey's divergence for the proportions. We then chose the 5,000 genes with the largest summed distance matrix. As described in Chapter 2.2.7, we performed consensus clustering (Monti et al., 2003) on top of each of our clustering routines in order to identify robust clusters. Briefly, this process involves repeated subsampling of the entire dataset and enumeration of how frequently samples were clustered together in a consensus matrix. After performing 1000 subsamples, a consensus matrix is calculated in which each entry is the fraction of times two samples were clustered together when they were sampled together. The final clustering is determined by performing clustering on the consensus matrix.

Additionally, consensus clustering is helpful in identifying outlier samples which do not align well with any other samples. In the process of our analysis, we identified 11 samples which did not cluster well with other samples at most K. Typically, these samples were found to have either very low overall expression or an overrepresentation of isoforms showing no expression. We removed these samples from further analysis as they were tending to drive clustering to several one-sample clusters rather than larger clusters.

4.2.3 Comparison in Identifying Batch Effect

In Figure 4.2 we compare the clustering assignments from clustering on the the three different features (gene, isoform, and proportion) in order to demonstrate the differences in how the samples are clustered. In this figure, we have also superimposed the "plate" from which the sample originated, which signifies the batch in which the samples were sent to be sequenced. We can see that in the cluster assignments of all of the methods, plate 734 clusters together. Given that this is public data from the TCGA project, it is difficult to know exactly what differences occur between plates, but they usually indicate batches of samples for which the mRNA extraction and other library preparation steps are done jointly.

In comparing the performance of the methods in detecting the plate artifact, it is clear that proportion clustering is almost perfectly detecting the plate effect, which is the primary signal driving clustering at $K = 2$. The other two methods do not have as clear a correspondence with plate. At $K = 2$, the isoform clustering method does capture some of the plate effect, though not as accurately as the proportion clustering method.

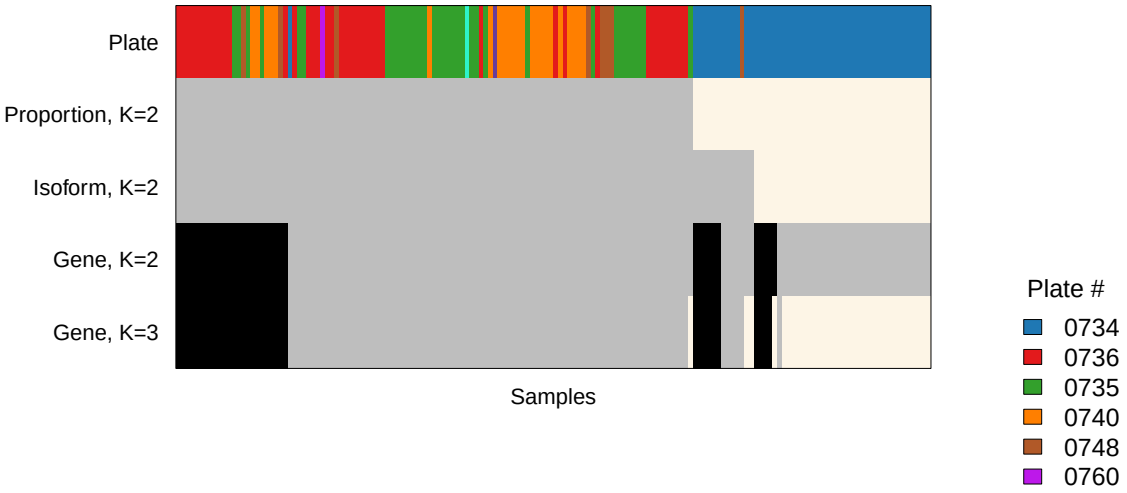


Figure 4.2: **Comparison of hierarchical clustering assignment.** Each row in this tracking plot corresponds to a clustering method and every column corresponds to an individual. The cluster assignments of the three different clusterings (isoform, gene and proportion) are shown for $K = 2$ groups by coloring the sample according to its clustering in the above plot. We also show $K = 3$ for gene clustering, since this is the point at which the gene clustering starts to have clusterings corresponding to the plate. The samples have been ordered to highlight the similarity between the clusterings. The top row shows the plate assignment of each sample.

At $K = 3$, the gene clustering method catches the signal with roughly the same accuracy as the isoform based clustering. However, this suggests that the batch effect is not the dominant signal being identified in gene expression clustering.

Similarly, the batch effect was also the dominant signal caught in proportion clustering when performed using K-medoid clustering, as seen in Figure 4.3. Again, isoform clustering also partially captures this signal, though not as accurately as proportion clustering. As expected, the gene clustering method does not identify the plate effect at $K = 2$. However, unlike in hierarchical clustering, gene clustering also does not capture the batch effect at $K = 3$.

One possible explanation for this clustering may be if a clinical variable were confounded with plate. In this case, it is possible that the clustering may be due to the clinical variable where all the individuals with that clinical variable ended up on the same plate. However, this was not the case here as individuals with certain clinical variables were distributed randomly among the different plates

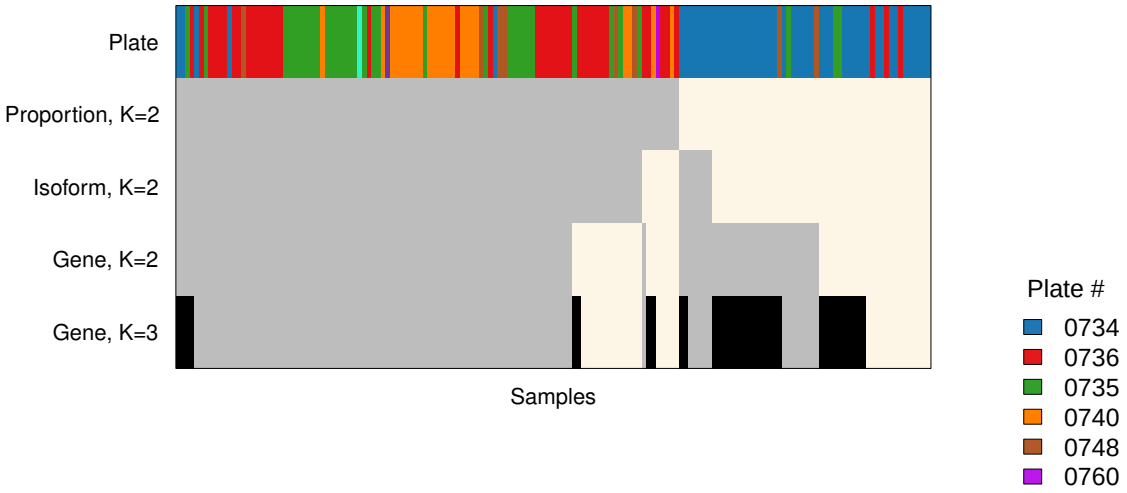


Figure 4.3: **Comparison of K-medoid clustering assignment.** Each row in this tracking plot corresponds to a clustering method and every column corresponds to an individual. The cluster assignments of the three different clusterings (isoform, gene and proportion) are shown for $K = 2$ groups by coloring the sample according to its clustering in the above plot. We also show $K = 3$ for gene clustering for comparison to Figure 4.2. The samples have been ordered to highlight the similarity between the clusterings. The top row shows the plate assignment of each sample.

4.2.4 Effects on Isoform Expression Due to Batch

As an examination of how prevalent the effect due to batch was, we looked at how many isoforms were differentially expressed between the two clusters found using the proportion based method at $K = 2$. Using **edgeR**, in all isoforms that showed expression above a certain mean expression level, we identified the isoforms which were differentially expressed between the two groups at a FDR of 0.05. The effect on isoform expression due to the batch effect was widespread, with 6,272 isoforms showing increased expression and 5,987 isoforms showing decreased expression in this plate. The expression patterns of the 2,500 most significant isoforms can be seen in Figure 4.4.

4.3 Re-analysis of Batch Effects in TCGA Data

Since the TCGA AML paper from which this data is drawn explicitly states that this data did not show a batch effect due to plate (Cancer Genome Atlas Research Network, 2013), we attempted to recreate the batch effect analysis of the TCGA working group as closely as possible. Specifically, the working group analyzed the level 3 data sets using both hierarchical clustering and Principal Components Analysis (PCA) to examine for plate

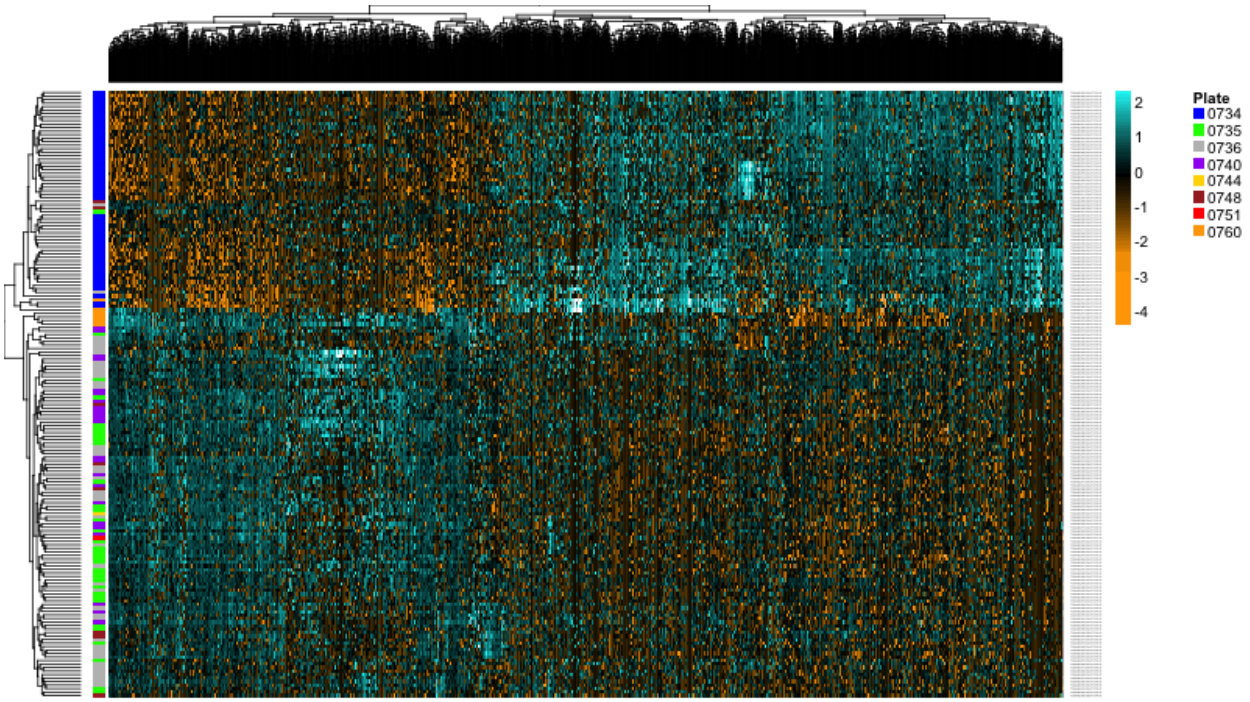


Figure 4.4: **Isoform expression for proportion clusterings showing batch effect.** Here we show a heatmap of isoforms found to be differentially expressed between the groups defined by hierarchical proportion clustering for $K = 2$. Each sample is a different row (colored by plate) with each isoform represented by a different column. The color scale represents whether an isoform is over-expressed or under-expressed relative to the mean level of expression for each isoform.

effects. We initially noted the presence of this batch effect in data we had processed using a Tophat and Cufflinks pipeline (Trapnell, Williams, et al., 2010), but in order to closely recreate the TCGA analysis, we retrieved the level 3 dataset directly from the TCGA data portal. Following the analysis described for hierarchical clustering, we used the Pearson correlation distance $(1 - r)$ as the dissimilarity measure. We then performed clustering utilizing the average linkage algorithm and annotated the dendrogram with colored bars denoting the different plates. For PCA, we plotted of the first two components and colored the patient id to denote the plate identity.

Since the TCGA batch effect analysis was performed on gene counts, we began by looking for batch effects in these counts. Beginning with all genes, we applied a low expression filter to the gene level counts, only keeping genes with a median greater than 25 counts across all individuals. In this case, the batch effect is clear in the PCA plot (Figure 4.5b), though not obvious in the hierarchical clustering plot (Figure 4.5a). When

we filtered the dataset to the 5,000 genes with the most variable counts, we noticed the batch effect was still present in this case, though not as obvious as when all the genes were used (Figure 4.5c). Though we were unable to find the exact number of genes that were examined in the TCGA batch effect analysis, other clustering analyses performed as part of the Firehose output were typically performed on 1,500 genes. We would expect that if the batch effect was not as noticeable in analysis performed on 5,000 genes as it was in analaysis performed on all the genes, it would only become less obvious as lower variable genes continue to be removed from analysis.

We then performed the same analysis on isoform values. Again, we see a strong batch effect when all isoforms are included (Figure 4.6a and Figure 4.6b), even more pronounced than we saw in the gene case. After we filtered down to the 5,000 most variable isoforms (Figure 4.6c and Figure 4.6d), we again see a reduction, but not total removal, of the effect.

4.3.1 Correcting Batch Effects

Since the batch effect is associated quite strongly with plate number, we were able to correct for it using `ComBat` in the **sva** package (W. E. Johnson, C. Li, and Rabinovic, 2007). In the model used by `ComBat`, the expression of a gene g is denoted by Y_{ijg}, where i indicates the batch and j indicates the sample. The expression may be modeled as

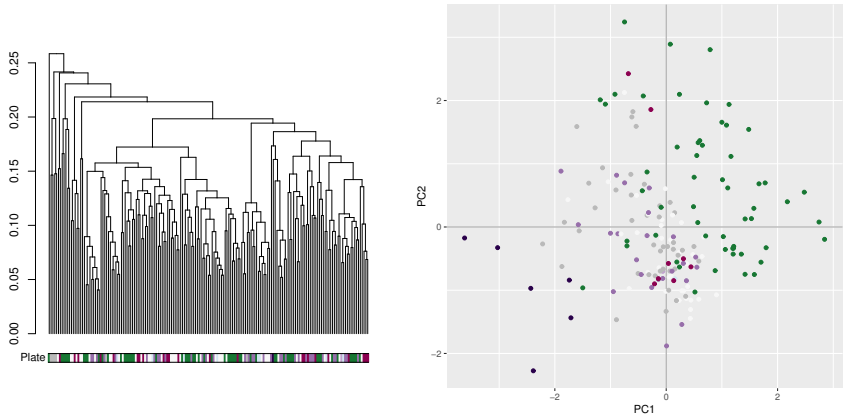
$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg},$$

where α_g is the overall gene expression, X is a design matrix for sample conditions, and β_g is the vector of regression coefficients corresponding to X. The error terms, ϵ_{ijg}, can be assumed to follow a Normal distribution with mean of zero and variance σ_g^2. The γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects of batch i for gene g, respectively.

After standardizing the data for each gene, batch effect parameters are estimated for each gene. After the estimation of individual parameters, Empirical Bayes methods are used to pool the information from all genes together in order to shrink the individual genes' batch effect parameter estimates towards the overall mean of the estimates. The pooled estimates may then be used to adjust the data for batch effects.

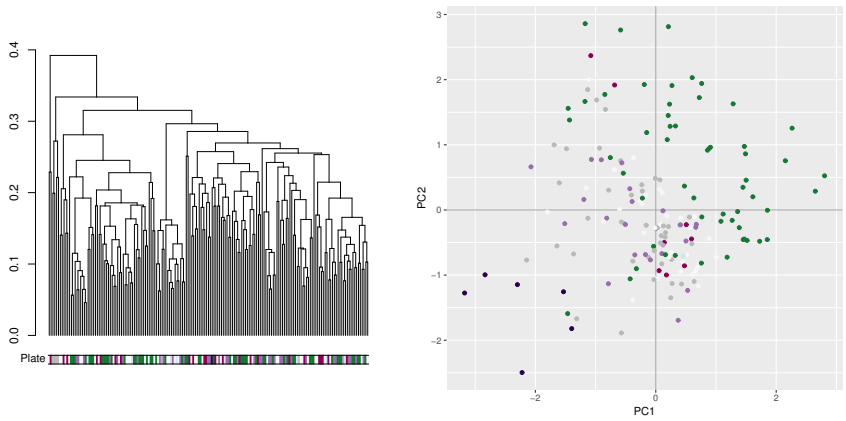
We performed `ComBat` on all isoform counts. Additionally, new gene counts were generated from the batch corrected isoform counts. The PCA plots and the hierarchical clustering shown in Figure 4.7 suggest that the batch effect has been removed from both isoform and gene counts.

Re-running the gene, isoform, and proportion clustering algorithms after removal of the plate effect with the batch correction tool `ComBat` (W. E. Johnson, C. Li, and Rabinovic, 2007) does result in different clusterings. It is useful to look at $K = 2$ for proportion clustering (Figure 4.8) as this was the clustering analysis which first identified the batch effect. By comparing the clustering before and after correction, we can see the samples are no longer clustering based on plate.



(a) Hierarchical Clustering (All Genes)

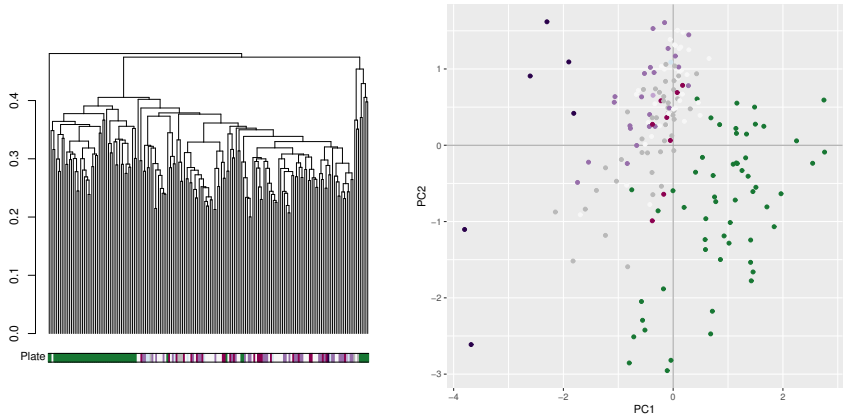
(b) PCA (All Genes)



(c) Hierarchical Clustering (Most Variable Genes)

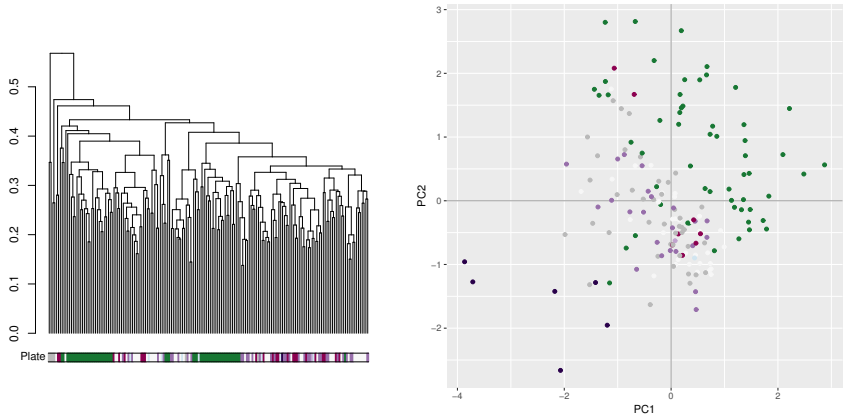
(d) PCA (Most Variable Genes)

Figure 4.5: **EDA of gene counts before batch correction (colored by plate).** Genes included in all analysis were filtered by median expression levels. Additionally, genes included in 4.5c and 4.5d were selected for analysis due to being among the 5,000 most variable genes in the TCGA LAML data set. Points in the PCA are colored by plate as is the colorbar beneath the dendrogram. The batch effect is more easily identified in the PCA plot than the dendrogram, though clustering by plate does occur in both.



(a) Hierarchical Clustering (All Isoforms)

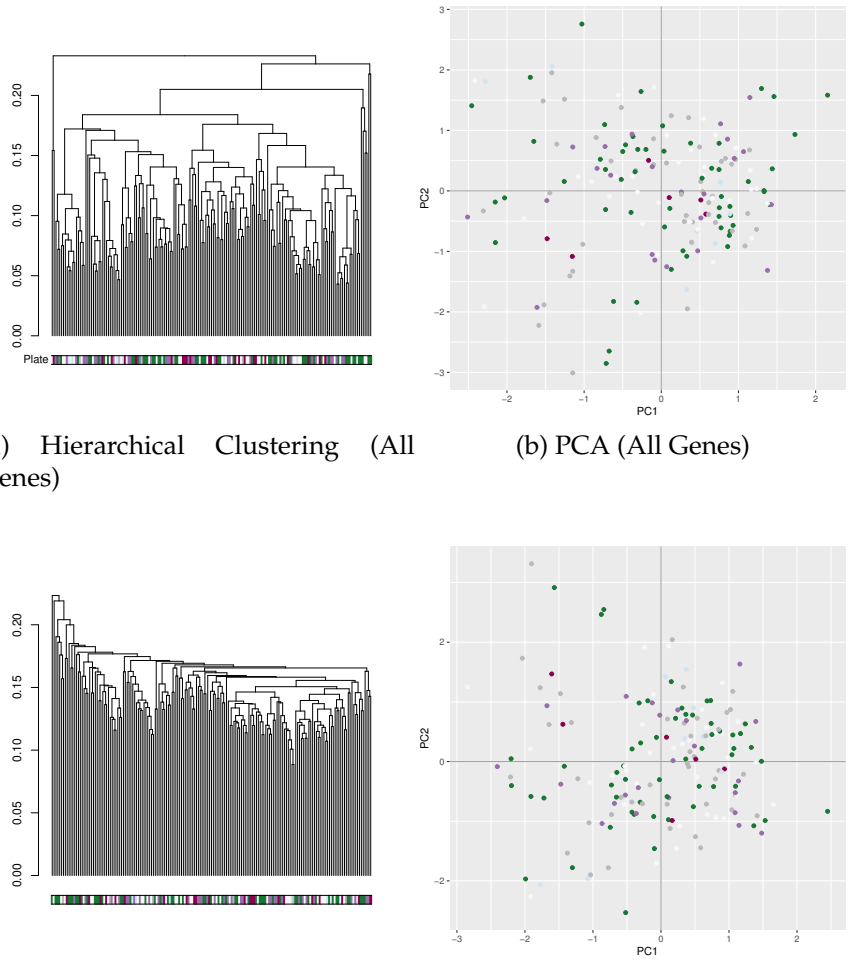
(b) PCA (All Isoforms)



(c) Hierarchical Clustering (Most Variable Isoforms)

(d) PCA (Most Variable Isoforms)

Figure 4.6: **EDA of isoform counts before batch correction (colored by plate).** Isoforms included in all analysis were filtered by median expression levels. Additionally, isoforms included in 4.6c and 4.6d were additionally chosen for analysis due to being among the 5,000 most variable isoforms in the TCGA LAML data set. Points in the PCA are colored by plate as is the colorbar beneath the dendrogram. The batch effect is readily apparent in all dendrograms and PCA plots.



(a) Hierarchical Clustering (All Genes)

(b) PCA (All Genes)

(c) Hierarchical Clustering (All Isoforms)

(d) PCA (All Isoforms)

Figure 4.7: **EDA of counts after batch correction (colored by plate).** Isoforms included in all analysis were filtered by median expression levels. Gene data after correction is seen in (4.7a) and (4.7b), while isoform data after correction is seen (4.7c) and (4.7d). Points in the PCA are colored by plate as well as the colorbar beneath the dendrogram. We no longer see separation based on plate ID.

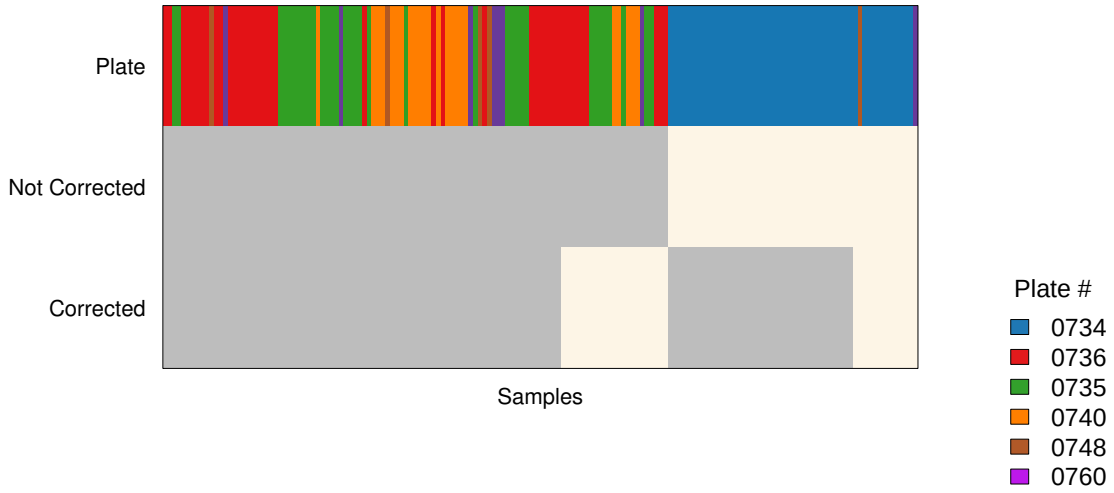


Figure 4.8: **Comparison of hierarchical clustering assignment before and after batch correction.** Each column corresponds to a sample and each row corresponds to clustering assignment determined by proportion clustering. The cluster assignments determined with or without batch effect correction are shown for $K = 2$ groups by coloring the sample according to its clustering in the above plot. The samples have been ordered to highlight the similarity between the clusterings.

Additionally, we can further look at gene clustering before and after correction to see the results of batch correction (Figure 4.9). Despite the fact that there was not an obvious batch effect in the gene counts when looking at the PCA plot or dendrogram of the counts, the gene cluster assignments do differ before and after correcting for the batch effect. This does suggest that the underlying batch effect did affect the gene clusterings obtained from the data. Additionally, we have included in this figure clinical variables which were found by the TCGA working group to be association with the cluster assignments (Cancer Genome Atlas Research Network, 2013). These clinical variables included somatic mutations in several genes (*U2AF1*, *TP53*, *NPM1*, and *RUNX1*), fusions (PML-RARA, MYH11-CBFB, RUNX1-RUNX1T1, and MLL Fusions), and the French-American-British (FAB) classification system. These clinical variables do correspond well to the cluster we determined. Interestingly, many of the clinical features also clustered together in the uncorrected clustering, suggesting that in gene clustering, the effect of these clinical variables was greater than the effect of the batch.

Another note from Figure 4.9 is that many of the patients with mutations, amplifications, and uniparetal disomy of *U2AF1* in the data set were clustered together. Interestingly, the patients with mutations in *U2AF1* are found in the same cluster as the patients with mutations in *TP53*. Initially, we were interested in *U2AF1* due to its involvement with splicing, and this suggests that potential effects of *U2AF1* mutation on isoform ex-

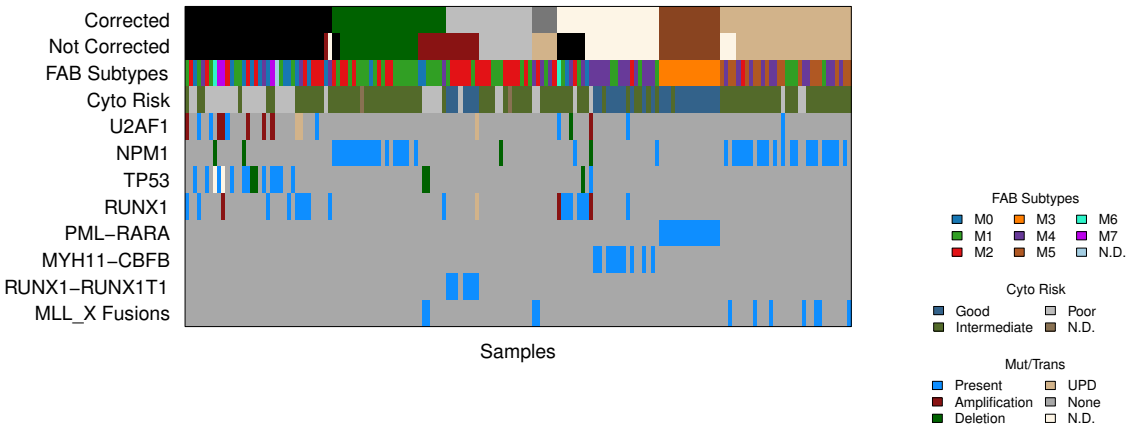


Figure 4.9: **Comparison of hierarchical clustering assignment before and after correction.** Each column corresponds to a sample and each row corresponds to either a clustering assignment or a clinical variable. The cluster assignments were performed by gene clustering with or without batch effect correction. Shown here are $K = 7$ groups denoted by coloring the sample according to its cluster. The samples have been ordered to highlight the similarity between the clusterings.

pression are also seen in gene expression.

4.3.2 Comparison of Gold Standard to Simulated Data

We were interested in comparing the data produced in our simulations to the TCGA LAML batch effect in order to evaluate how reasonable the simulations were. In order to assess this, we wanted to compare the difference in the two groups being clustered in real and simulated data.

In order to measure how well formed clusters are, we can look at silhouette scores. The silhouette score describes how similar an individual is to the members of its cluster compared to other clusters present. In order to calculate the silhouette score, we need to know the value $a(i)$, which is the average dissimilarity between an individual i and all of the other members of its cluster. Additionally, we need to know the value $b(i)$, which is the average dissimilarity of individual i to the members in the cluster that is the next best fit for i. The silhouette score can be calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

We can see that $s(i)$ will vary between -1 and 1. If the value is close to 1, then i has been cluster appropriately. If the value is close to -1, then i is more likely the next best fit

cluster than the one i is currently in. A value of 0 suggests that either cluster would be a reasonable fit.

To compare our simulation clusters to the batch effect clusters, we wanted to generate simulated data for $K = 2$. In order to do this, we simulated 7,501 genes in which 10% of the genes contained a true cluster at $K = 2$. For each gene, we then found the mean of the silhouette scores across all individuals in that gene. When we considered only the genes with the pattern, the mean silhouette score was 0.58. We can see in the boxplots in Figure 4.10 that there is a clear distinction between the distribution of the silhouette scores of both simulated groups. For the simulated data, the overall mean of the silhouette scores was 0.057, which suggests little preference for a cluster. In only the genes with noise, the mean was estimated to be -0.0012. The simulated data only contained 10% of the genes with the pattern so the vast majority of silhouette scores would have been for genes containing noise. This suggests an average silhouette score near 0 to be reasonable for the simulated data.

In the real data, we calculated the mean of the silhouette scores across all the genes to be 0.19. Since we do not know which genes were actually affected by the batch effect in the real data, we can not break the genes down into genes with the pattern and genes without the pattern. We note that the average silhouette scores in the real data are not as high as those seen in the simulated data with groups. That is, on average, the clusters in the simulated groups are more dissimilar that those seen in the real data. However, Based on the silhouette averages as well as the boxplots in Figure 4.10, we assume that the batch effect affects a much larger percentage of genes than the 10% we simulated in this data.

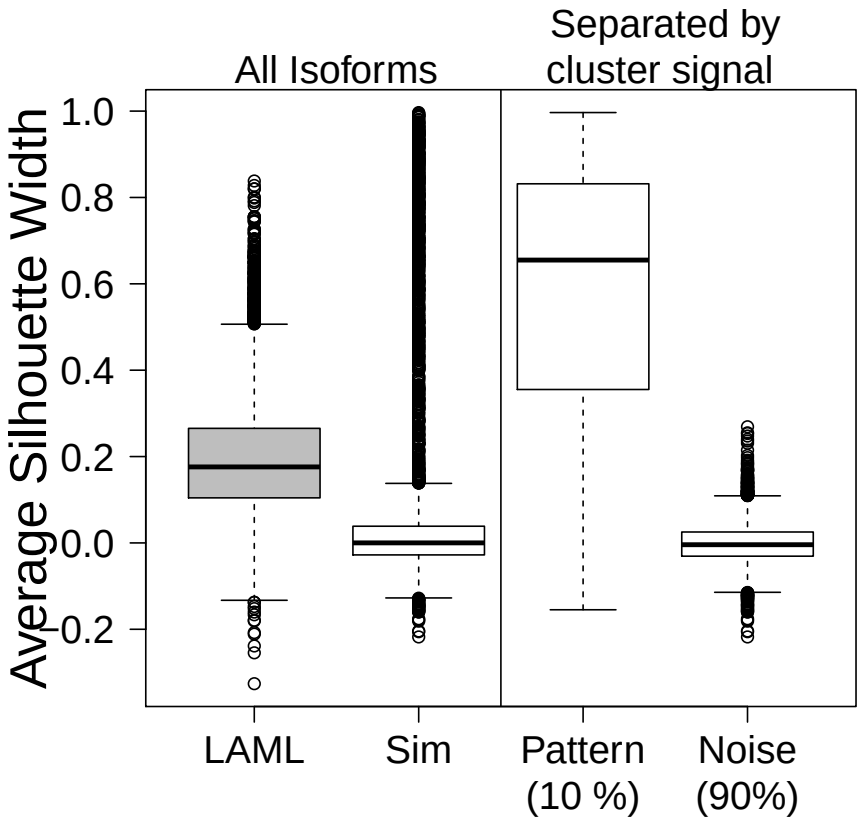


Figure 4.10: **Comparison of silhouette scores in simulated and real data.** The right panel shows the silhouette scores of our simulated data grouped by whether or not they were simulated to contain a difference in relative isoform usage. The left panel shows all silhouette scores from the TCGA LAML data set as well as all silhouette scores from the simulated data. The silhouette scores from the LAML data set have a much higher median, which is expected as the batch effect appears to effect more than 10% of the data.

Chapter 5

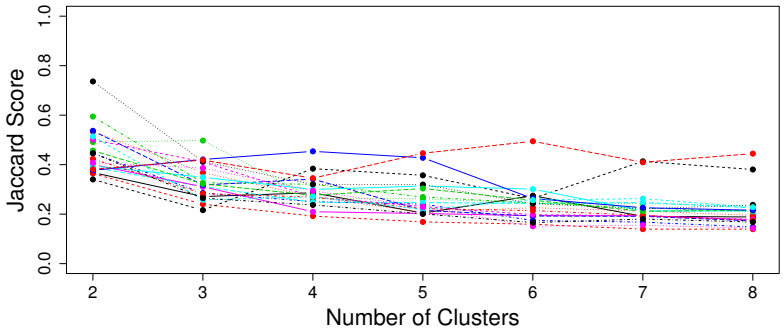
Real Data: Identifying Subtypes

5.1 Pan-Cancer Analysis

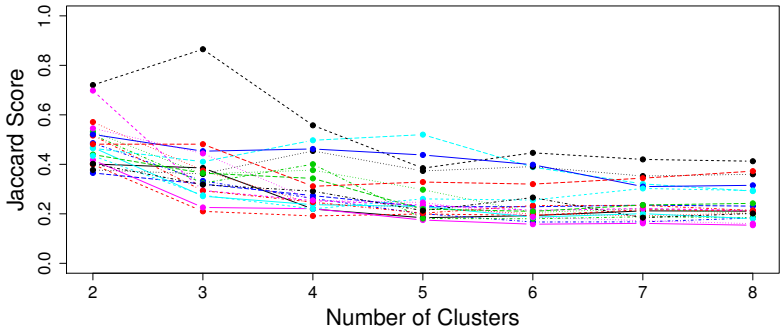
5.1.1 Similarity Between Cluster Assignments

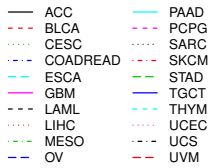
We ran the gene, isoform, and proportion clustering algorithms on several TCGA data sets including adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon and rectum adenocarcinoma (COADREAD), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), liver hepatocellular carcinoma (LIHC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), sarcoma (SARC), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS), and uveal melanoma (UVM). For each cancer type, we ran the gene cluster algorithm on the 5,000 most variable genes. We ran the isoform clustering algorithm on the 5,000 most variable isoforms, selected from genes which had multiple isoforms above a certain expression level. We ran the proportion clustering algorithm on the 5,000 genes which had the highest mean distance matrices after calculating each distance matrix based on distance between relative isoform frequencies using Jeffrey divergence as the proportion distance measure. Similar to the isoform clustering case, these relative isoform frequencies were calculated after filtering out isoforms with low expression levels.

One question we were interested in exploring with these various data sets was whether these different clustering algorithms would return different results for gene, isoform, and proportion based clustering methods. Despite the high number of isoforms we saw in our data sets, we wondered if most of the changes seen in gene expression were actually due to changes in only one major isoform per gene. One group which saw little change in relative isoform frequency postulated that 60% of the variability seen in isoform levels is actually due to changes in gene expression rather than changes in splicing ratio (Gonzàlez-Porta et al., 2012). If this were the case for most genes, we would expect that



(a) K-medoids





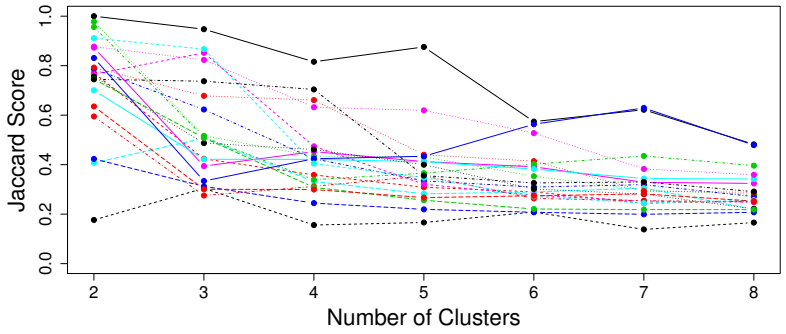
(b) Hierarchical

Figure 5.1: **Similarity between gene and isoform in K-medoids and hierarchical clustering:** We see in this comparison over 20 TCGA data sets that cluster assignments derived from gene and isoform clustering have relatively low similarity, which generally decreases with increasing K.

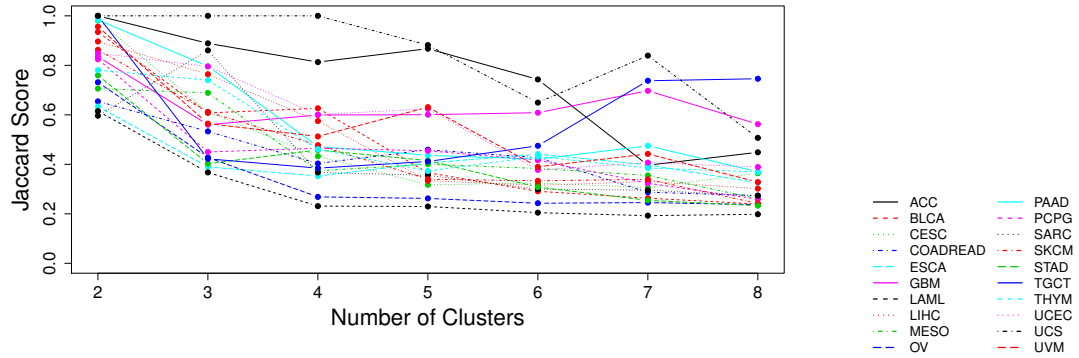
gene clustering and isoform clustering would be very similar. To examine the similarity between the cluster assignments, we calculated the Jaccard score between cluster assignments for K varying from 2 to 8.

In comparing gene and isoform clustering (Figure 5.1), we note that many of the different data sets have a relatively low similarity score even at $K = 2$, typically no more than 0.6 for most tumors. Also, for most tumors, the Jaccard score tends to decrease with increasing K, suggesting that clusters generally become less similar as the number of clusters increases. One notable exception is found in the hierarchical clustering results for LAML (Figure 5.1b), which has a score around 0.83 at $K = 3$, which is where the similarity between gene and isoform clustering peaked for that data set.

In comparing proportion and isoform clustering (Figure 5.2), we note that these Jac-



(a) K-medoids



(b) Hierarchical

Figure 5.2: **Similarity between proportion and isoform clustering in K-medoids and hierarchical clustering:** We see in this comparison over 20 TCGA data sets that cluster assignments derived from isoform and proportion clustering are generally more similar that seen in isoform and gene clustering (Fig 5.1). Some data sets, such as ACC, UCS, and TCGT, show relatively high Jaccard scores even with increasing K.

card scores are much higher than those seen in the comparison between gene and isoform clustering. At $K = 2$, the Jaccard scores for most cancer are above 0.6, with a few examples of scores of 1, a perfect match between the cluster assignments. The UCS data set has perfect concordance between isoform and proportion clustering at $K = 2$, $K = 3$, and $K = 4$. As was the case between gene and isoform clustering, the similarity between proportion and isoform clustering seems to decrease with increasing K. An exception to this is the TGCT data set, which shows quite high similarity between proportion and isoform clustering in hierarchical clustering even at $K = 8$.

Comparison of these clustering results suggest that the gene, isoform, and proportion clustering algorithms will give different clustering results. This is particularly the case

for gene and isoform clustering in the data sets examined. One reason for this is that genes which only showed expression in one isoform were not included in the isoform clustering data set. Additionally, this may also suggest that genes with multiple isoforms did not necessarily always have one major isoform that dominated expression. Proportion and isoform clustering had higher similarity scores than those seen between gene and isoform. Unlike gene clustering, we started with a similar data set for these two algorithms. Though different filters were applied later in each algorithm, both proportion and isoform clustering algorithms started with the set of genes containing multiple isoforms above a certain expression value.

5.1.2 Association with Clinical Variables

As discussed in an earlier chapter, some cancer specific alternative splicing may occur as isoforms that are typically silenced in normal tissue are instead expressed in tumor tissue. This may particularly be true for isoforms which promote growth, proliferation, or other traits that are beneficial to cancer. We were interested in whether the cluster assignments found using our clustering methods might be associated with different stages of tumor growth.

The clinical information available for many of the TCGA data sets included staging information for each patient's cancer. Staging is a way of describing the size of a tumor and how far it has grown. The stage is based on four factors: location of the original tumor, size of the tumor, lymph node involvement, and presence of metastatis. The two methods of staging most cancers are the number staging system and the TNM system. The T category refers to the size of the tumor. This value varies from 1 to 4, with the value increasing with increasing tumor size. The N category refers to whether the cancer has spread to the lymph nodes and ranges from 0 to 3, with the value increasing as the number of affected lymph nodes increase. The M category refers to whether the cancer has metastasized to another part of the body with a value of 1 if the cancer has metastasized and 0 otherwise. Number staging systems usually use the TNM system to divide cancers into stages.

Based on the staging information available, we looked at whether there was an association between the different cancer stages and isoform or proportion cluster assignments for the following cancer types: ACC, BLCA, CESC, COADREAD, ESCA, LIHC, MESO, OV, PAAD, SKCM, STAD, TGCT, THYM, UCEC, UCS, and UVM. We performed a χ^2 test of independence to look for association between cluster assignments and cancer staging. After correcting for multiple hypothesis testing, we saw no significant association between the isoform and proportion cluster assignments and cancer stages in these data sets.

5.2 Acute Myeloid Leukemia

5.2.1 Comparison to Clinical Variables

We were particularly interested in the LAML data set as TCGA exome sequencing results had indicated that there were several mutations present in a gene encoding a known splicing factor, *U2AF1* (Cancer Genome Atlas Research Network, 2013). In the previous chapter, we described our work with clustering this data set and presented our identification of an unreported batch effect in that data. After correcting for the batch effect using `ComBat` in the **sva** package, we reanalyzed the data set with our clustering algorithms. Figure 5.3 shows the resulting cluster assignments at $K = 7$, to correspond with the K used in the TCGA LAML paper for gene clustering (Cancer Genome Atlas Research Network, 2013).

One thing to note in this figure is the difference between the three methods at $K = 7$. At this K, the methods with the most similar results are gene and isoform clustering. This suggests somewhat similar patterns in gene and isoform expression, which are driven by the clinical variables included here that were mentioned by TCGA as being associated with the gene clustering assignments (Cancer Genome Atlas Research Network, 2013). The proportion clustering results are quite different than the results from these two methods. Many of the clinical variables included are not associated strongly with the proportion clustering groups. For example, patients with *NPM1* mutations were found in only two gene clustering group, while they are spread throughout five proportion clustering groups. Similarly, the patients with PML-RARA, MYH11-CBFB, and RUNX1-RUNX1T1 fusions were contained within one cluster in gene clustering but are distributed in multiple clusters in proportion clustering.

However, we do note that mutations and copy number variation of *U2AF1* may be driving the clustering of at least one of the groups in proportion clustering, the dark gray cluster in Figure 5.3. We note that in this cluster of 14 patients, four patients have mutations in *U2AF1*, four patients have amplication of *U2AF1*, and one patient has uniparental disomy of *U2AF1*. Additionally, another cluster of 10 patients (the dark brown group) contains one with *U2AF1* mutation, one with amplification of *U2AF1*, and one with uniparental disomy of *U2AF1*. A χ^2 test of association returned a p-value of 3.3×10^{-7}. Additionally, at $K = 6$, most members of these two clusters were clustered together in the proportion based clustering (see Figure 5.4). At $K = 6$, a group of 26 patients contained five patients with mutations, six patients with amplification, and one with uniparental disomy of *U2AF1*. A χ^2 test of association returned a p-value of 2.5×10^{-6}. However, we also note that gene clustering (see Figure 4.9) also shows similar clustering of many of these patients. Using the gene clustering results for $K = 7$, one cluster of 36 invididuals contained four patients with mutations, six patients with amplification, and two with uniparental disomy of *U2AF1*. Several groups had found that mutations in *U2AF1* resulted in differential splicing of hundreds of genes (Ilagan et al., 2015; Przychodzen et al., 2013), though it is unclear if amplification of *U2AF1* would have the same affect on differential

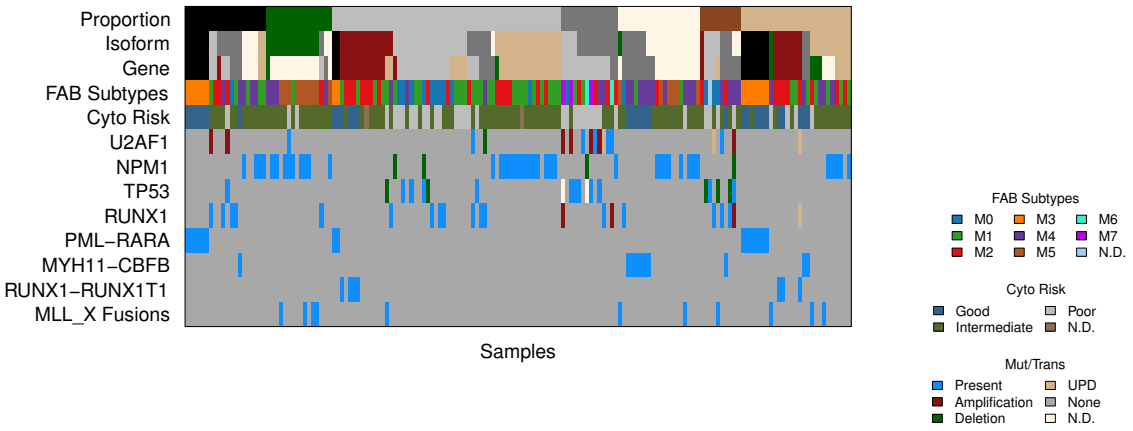


Figure 5.3: **Comparison of hierarchical clustering assignment to clinical data.** Each column corresponds to a sample and each row corresponds to either a clustering assignment or a clinical variable. The cluster assignments were performed by proportion, isoform, and gene clustering with batch effect correction. Shown here are $K = 7$ groups denoted by coloring the sample according to its clustering in the above figure. The samples have been ordered to highlight the similarity between the proportion clusterings.

isoform usage. The fact that these individuals cluster together in both gene clustering and proportion clustering suggests that these clusters may be the result of expression changes in a single major isoform rather than an event such as isoform switching.

Additionally, we see in Figure 5.4, not only do patients with mutations and copy number variants of *U2AF1* seem to cluster together, but these same clusters also contain patients with mutations and deletions of *TP53* and mutations and amplifications of *RUNX1*. In fact, four patients had both a mutation in *TP53* and an amplification of *U2AF1*. As a result of these confounding variables, it makes it difficult to say that mutation is *U2AF1* are driving this clustering.

5.2.2 Comparing Proportion and Isoform Clustering

Visually, we saw that the isoform clustering for LAML looked more similar to gene clustering than proportion clustering. By using the Jaccard score as a measure of similarity, we can see in Figure 5.5 that the gene and isoform clustering results have a higher Jaccard score across all K than isoform clustering with proportion clustering in both hierarchical and K-medoid clustering. However, since exact cluster matches would result in a Jaccard score of 1, the cluster results between gene and isoform clustering are not exact. At small K when $K = 2, 3,$ or 4, it is also worth noting that the clusters obtained from hierarchical clustering are more similar than those obtain from K-medoid clustering. At higher K, the difference between the similarities calculated from hierarchical and K-medoid clustering

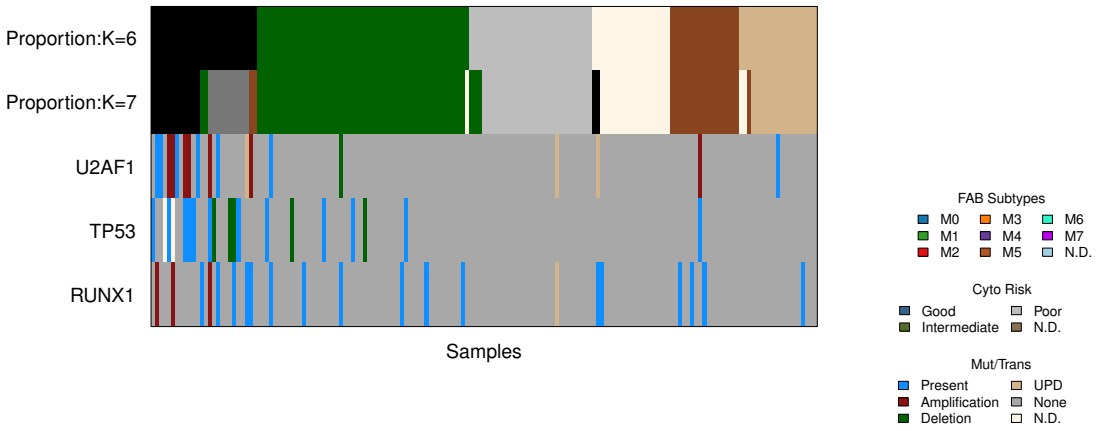


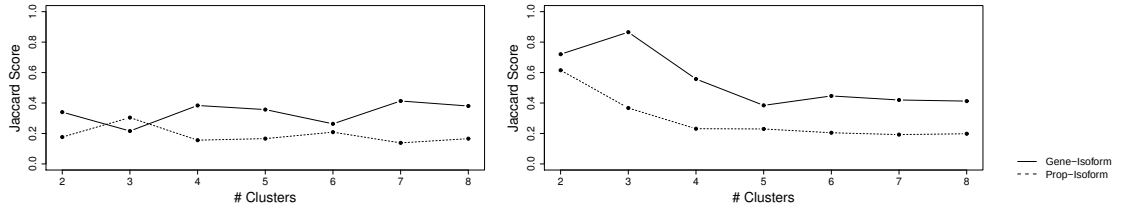
Figure 5.4: **Comparison of hierarchical proportion clustering assignment to *U2AF1* Data.** Each column corresponds to a sample and each row corresponds to either a clustering assignment or a clinical variable. The samples have been ordered by cluster assignment. We see some clustering of the samples with mutations and amplifications in *U2AF1*. However, other clinical variables also cluster within these assignments, including *TP53* mutations.

is small.

5.2.2.1 Comparison at $K = 2$

Since the difference in results between the three clustering algorithms occurs even at small K, we can look at the clustering result at $K = 2$ as the cases at smaller K are simpler to examine. Changes in relative isoform frequency may be the result of one isoform being differentially expressed between two groups while the other isoforms in the gene are relatively constant. In another case, relative isoform frequency may change due to contrasting expression pattern. That is, multiple isoforms show differential expression between groups, with one isoform undergoing up-regulation while another isoform undergoes down-regulation. This type of pattern would not be easily capture unless the relationship between isoforms were considered in the distance measure. As a result, we expect the clusters found in proportion clustering to show enrichment of genes with isoforms showing this type of switching.

In order to compare the clustering results for isoform and proportion clustering, we first determined which isoforms were clustered well by each method. Using all isoforms that showed expression above a certain mean expression level, we used the **edgeR** package to identify the isoforms that were differentially expressed between groups in the $K = 2$ clustering for the two clustering methods. We selected all isoforms that showed differential expression between the two clusters in each method at a FDR of 0.05. A



(a) K-medoid Clustering (b) Hierarchical Clustering

Figure 5.5: **Comparison of clusting assignments in TCGA LAML data set:** By calculating the Jaccard score, we can look at the similarity of gene and isoform clustering as well as isoform and proportion cluster. We see that cluster assignments for gene and isoform clustering using hierarchical clustering are similar at several K, including $K = 2$ and $K = 3$. The isoform and proportion cluster assignments using hierarchical clustering are not as similar, nor are any cluster assignments found using K-medoid clustering.

heatmap of the most significant isoforms is shown in Figure 5.6.

For each method, we further broke down the differentially expressed isoforms based on whether the log fold change of that isoform was positive or negative. At a FDR of 0.05, we found that 13,656 isoforms showed significant mean expression difference between the two groups in the isoform based clustering method. Of these 6,427 isoforms had expression similar to the I2 (purple) group, while 7,229 had expression similar to the I1 (green) group. We also found 12,526 isoforms showing significant difference between the groups found using the proportion based clustering method. We determined that 6,207 isoforms showed expression similar to the P1 (red) group, while 6,319 showed a pattern similar to the P2 (blue) group. If we inspect the isoforms that constitute the I1 and I2 isoform group (those whose expression patterns correspond highly with the isoform), it was rare for isoforms from a single gene to be both in I1 and I2 (Table 5.1). When compared to the genes whose isoforms constitute the P1 and P2 isoform groups (those whose expression patterns correspond highly with the proportion), we note these groups contain a higher proportion of genes with an isoform in both P1 and P2. This suggests that the proportion clustering is detecting isoform switching patterns within a gene, and that such patterns may not as easily be detected by clustering directly on isoforms due to the presence of other expression patterns that dominate the signal.

In examining genes that show this switching pattern, we found several genes with known relationship with AML or other cancers, including *SON*, *RBBP6*, *SENP5*, and *CHD8*. As one example, the gene *SON* encodes a splicing factor involved in regulation of the machinery of splicing. The expressed isoforms of *SON* are seen in Figure 5.7. Note that different sets of isoforms were found to be differentially expressed in the two different clustering methods. Recently it has been suggested that different isoforms of the *SON* gene result in different regulation of the mixed lineage leukemia gene (MLL) com-

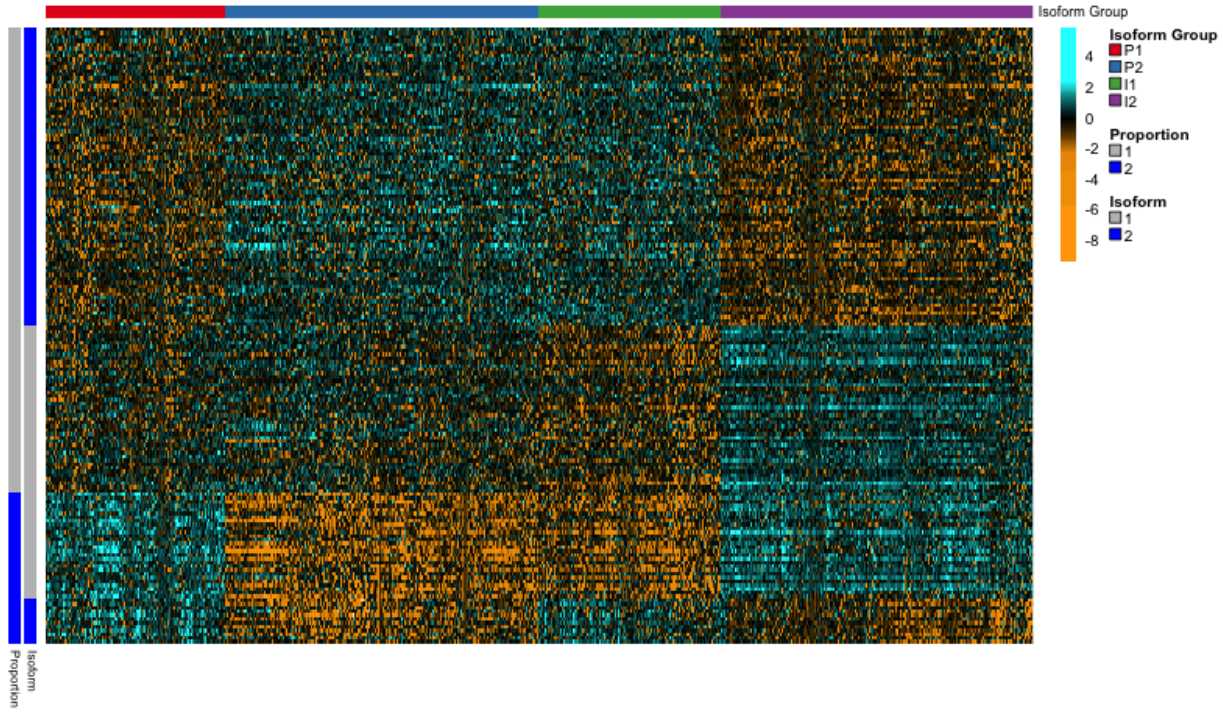


Figure 5.6: **Isoform expression for proportion and isoform clusterings** Here we show a heatmap of the isoform expression found to be differentially expressed either between the proportion clustering groups or the isoform clustering groups for $K = 2$. The individual samples are denoted by the rows and the columns are individual isoforms. The color scale represents whether an isoform is over-expressed or under-expressed relative to the mean level of expression for each isoform. To the left of the heatmap, a separate color scale identifies the samples in the proportion and isoform clustering groups. Along the top of the heatmap are assignments of isoforms to different groups of isoforms, for referencing in the text.

Overlap Groups	Groups P1	P2	I1	I2
P1	3882	561	–	–
P2	561	3241	–	–
I1	–	–	3508	292
I2	–	–	292	4210

(a) Number of genes

Overlap Groups	Groups P1	P2	I1	I2
P1	–	0.173	–	–
P2	0.145	–	–	–
I1	–	–	–	0.069
I2	–	–	0.083	–

(b) Proportion of genes

Table 5.1: The proportion of genes sharing isoforms between groups P1 and P2 is higher than between groups I1 and I2

plex assembly that activates several leukemia-associated target genes (Kim et al., 2016; Yokoyama et al., 2005). While we cannot assume this is the phenomena we detected with only this limited knowledge, these examples at least provide some support to the idea that the clustering we find with proportion clustering is finding isoform switching behavior within genes that could give meaningful biological results. The example of the splicing factor SON also demonstrates to how widespread patterns in alternative splicing could occur, since different isoforms of *SON* also regulate splicing.

5.2.2.2 Comparison at higher K

For $K = 2$, we described how the proportion based clustering method resulted in clusters that were enriched for genes with shared isoforms between clusters. We wanted to check if this was also true at higher K. At each K, we found the percentage of genes that had isoforms with contrasting expression patterns in the identified clusters. That is, we found the cases in which one isoform showed significant increased expression in a group while another isoform from the same gene showed decreased expression in that group. Specifically, we examined $K = 3$ (Table 5.2), $K = 4$ (Table 5.3), and $K = 5$ (Table 5.4). These Tables may be found at the end of this chapter.

It is noteworthy that we continued to see the pattern we noticed at $K = 2$. That is, we saw enrichment of contrasting isoforms in clusters formed by performing proportion based clustering. However, we additionally see that genes having this switching pattern remained a relatively low percentage of the genes with differentially expressed isoforms.

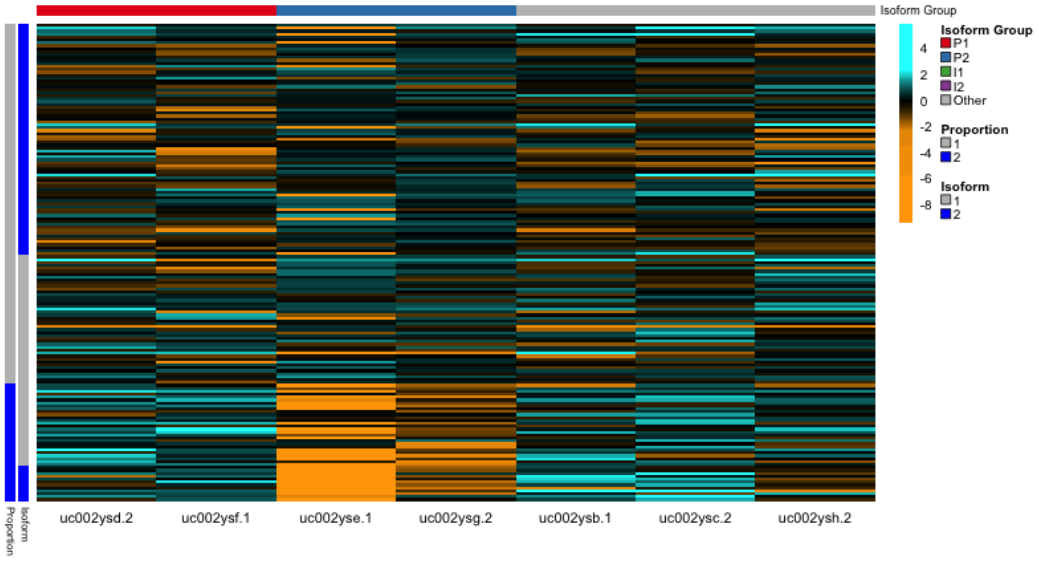


Figure 5.7: **Isoform expression of *SON* gene for proportion and isoform clusterings** Here we show a heatmap of the isoform expression for the *SON* gene. Row denote the samples while columns denote the individual isoforms. The color scale represents whether an isoform is over-expressed or under-expressed relative to the mean level of expression for each isoform. To the left of the heatmap, a color scale gives the identification of the samples to the proportion and isoform clustering groups. Along the top of the heatmap are assignments of isoforms to different groups of isoforms, for referencing in the text.

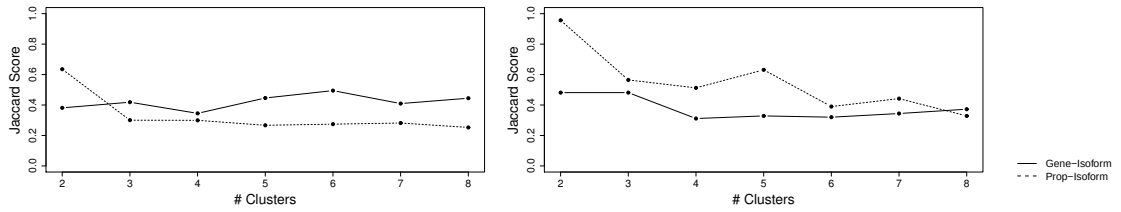
This suggests that the presence of genes with contrasting isoforms differentially expressed is uncommon. We see that many genes with only one differentially expressed isoform are driving clusters being identified even by the proportion based method.

5.3 Uveal Melanoma

5.3.1 Comparing Proportion and Isoform Clustering

We were also interested in studying the splicing of UVM tumors as the splicing factor *SF3B1* has been found to have mutations in around 15-20% of uveal melanoma cases and 10% of chronic lymphocytic leukemia tumors (Furney et al., 2013; Gentien et al., 2014; Quesada et al., 2012). Specifically, in the TCGA UVM data set, mutations in *SF3B1* were found in 18 out of 80 patients (22.5%).

Using the same methodology as described for the LAML and other TCGA data sets,



(a) K-medoid Clustering (b) Hierarchical Clustering

Figure 5.8: **Comparison of clusting assignments in TCGA UVM data set:** By calculating the Jaccard score, we can look at the similarity of gene and isoform clustering as well as isoform and proportion cluster. We see that hierarchical cluster assignments for proportion and isoform clustering are quite similar at $K = 2$. The isoform and gene cluster assignents were not as similar in either K-medoid or hierarchical clustering.

we also performed gene, isoform, and proportion based clustering on the UVM data set. After filtering isoforms based on very low expression, our data had 7,009 genes with multiple isoforms, consisting of 20,514 isoforms. Including single isoform genes, 13,507 genes had expression levels above a certain cutoff. Using the Jaccard score, we compared the similarity between isoform and gene clustering as well as isoform and proportion clustering (see Figure 5.8). For hierarchical clustering, we note that at $K = 2$, proportion and isoform clustering show a very high concordance, having a Jaccard score of 0.96. The Jaccard score for gene and isoform clustering for the same conditions is 0.48.

At $K = 2$, isoform and proportion clustering are identifying a similar signal that is not being found in gene clustering. This suggests that the isoform and proportion clustering methods may both be catching isoform switching events which do not result in large changes in gene expression differences. In fact, we see that 1053 of genes have both a differentially upregulated and downregulated isoform. This is 28.2% of all genes with an upregulated isoform (3,570 genes) and 29.5% of all genes with a downregulated isoform. Further, when we look at which genes show differential expression in this same cluster, only 619 of the 1053 genes (58.8%) showing contrasting isoforms show differential expression in gene counts. This suggests that in some cases, summarizing by genes expression rather than isoform expression loses some of the underlying isoform changes which occur. One such example is shown in Figure 5.9. We see in this example a dramatic shift in the preferred isoform expressed in each cluster. However, the gene expression level does not capture any of this signal.

5.3.2 Comparison over many K

We examined how the clustering varied depending on K in proportion clustering (Figure 5.10 and noticed there were three major clusters. These are denoted as the light gray

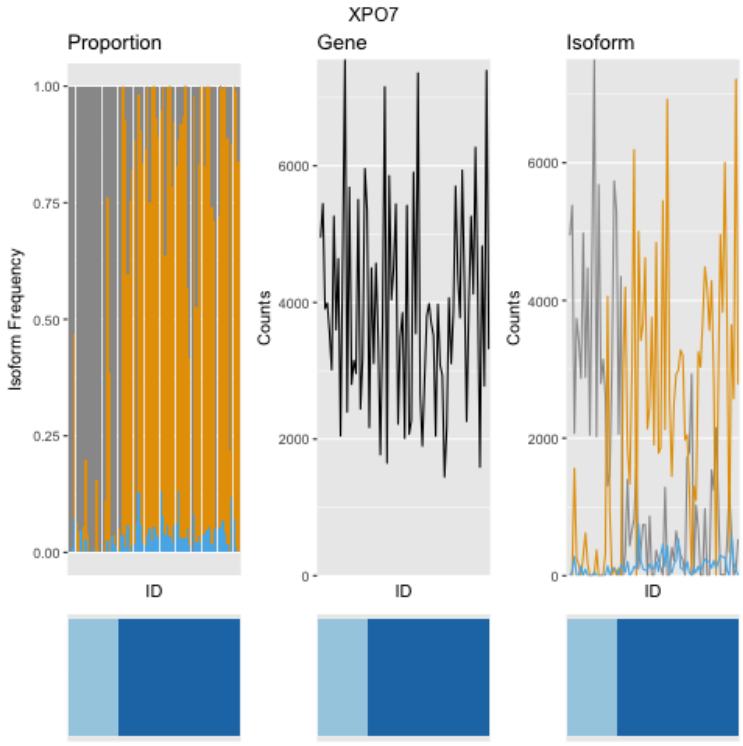


Figure 5.9: **Relative isoform frequency, gene, and isoform Levels of XPO7** Here we show gene and isoform expression as well as relative isoform frequency of XPO7, a protein involved in nuclear export of proteins. In the left most figure, we plotted the relative isoform frequency of each isoform. The x-axis is each individual and each isoform is represented by a different color. The proportion of different colors in the columns denotes the relative frequency of each isoform. The middle figure shows gene expression, and the x-axis is each individual. The right most figure shows isoform expression. Again, the x-axis is each individual, and each isoform is represented by a different color. We note the isoform that is preferentially expressed switches dramatically, though this does not manifest as a noticeable change in gene expression.

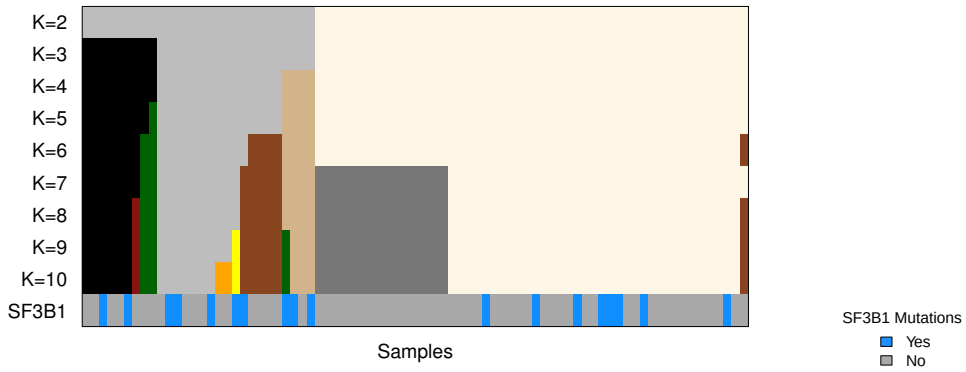


Figure 5.10: **Comparison of hierarchical clustering assignment to mutation data.** Each column corresponds to a sample and each row corresponds to either a clustering assignment or the presence or absence of a *SF3B1* mutation. The cluster assignments shown are clustering based on proportions only from $K = 2$ to $K = 10$ groups. The beige and dark gray clusters change little with increasing K, suggesting these are robust, stable clusters.

cluster at $K = 2$ and the dark gray and beige clusters at $K = 7$. We note the light gray cluster is the same as the light blue cluster in Figure 5.9. The dark green and beige clusters (which together make the dark blue cluster in Figure 5.9) remain fairly stable, while the light gray cluster becomes more and more fragmented at higher K.

We also noticed that in the three main clusters, one cluster contained no patients with *SF3B1* mutations (the dark gray cluster) with the patients with a mutation in *SF3B1* split equally between the two remaining clusters. We further looked at the dendrogram produced by the proportion based distance (Figure 5.11). We noticed that the patients with *SF3B1* mutations in the beige group clustered strongly together in the dendrogram produced by the proportion clustering method. In the light gray cluster, we see some patients who cluster with other patients with mutations, but not as tightly as the *SF3B1* mutation patients have clustered in the beige group.

We identified a strong differential isoform usage signal between the groups at $K = 2$, and it is perhaps not surprising that the changes due to the splicing mutation did not cause a signal stronger than this initial cluster breakpoint. Focusing only on the samples in the dark gray and beige clusters, we found which isoforms were differentially expressed between the patients with and without *SF3B1* mutations. We determined 1,157 isoforms from 892 genes were differentially expressed between these clusters. This is only 25% of the number of genes that had at least one differentially expressed isoform using the clusters found at $K = 2$. Not suprisingly, many of the genes with isoforms showing differential expression patterns in the presence of *SF3B1* mutations have been described before, including genes such as *ABCC5*, *UQCC* (see Figure 5.12), *ANKHD1* Furney et al.

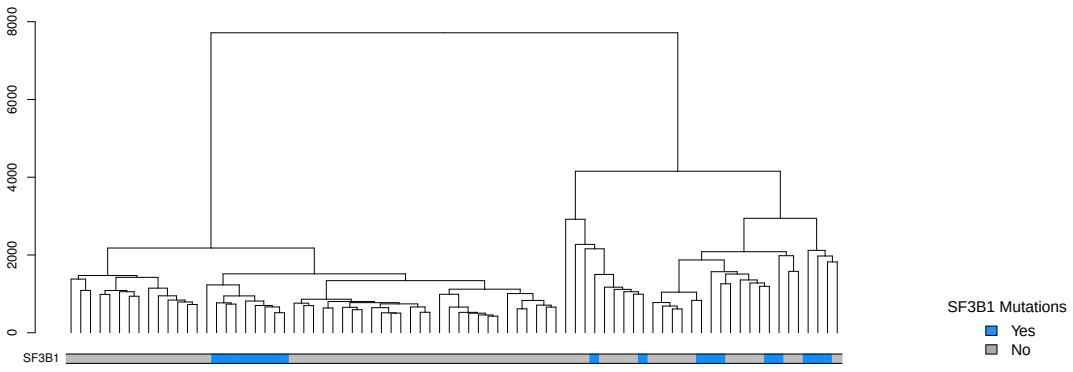


Figure 5.11: **Dendrogram of hierarchical clustering** This dendrogram shows the hierarchical clustering of the UVM data based on measuring distance using relative isoform frequency. The colored bar denotes the presence (in blue) of a *SF3B1* mutation.

(2013).

5.4 Conclusion

We proposed a novel method of clustering mRNA sequencing data in order to incorporate isoform information. In the case of mRNA sequencing data, the feature we are using to cluster are individual genes represented by relative isoform frequencies rather than counts. We presented simulations in which clustering on isoform proportion usage was more accurate at finding clusters based on differential isoform usage than using isoform counts alone. Additionally, we adapted a method of sparse clustering in order to select the most relevant features. However, this performed worse on the simulations than applying a simple variance filter to the data.

We used this method to identify a batch effect in the TCGA LAML data set. We showed that this data set had a 5′-3′ bias, which was manifested as a difference in coverage at the 5′ end versus the 3′ end for different plates. Relative isoform usage was most accurate at identifying which individuals had been processed on the different plate compared to clustering on isoform or gene counts alone.

In addition, we ran our algorithm on a number of different TCGA data sets. While we saw no association with relative isoform frequency clustering and tumor staging, we saw some potential examples of clustering due to mutations in splicing factors. In clustering results from the TCGA LAML data set, one cluster found from clustering on relative isoform frequency contained most of the individuals with mutations and amplifications in *U2AF1*. However, several other genes also showed a high rate of mutation in that cluster,

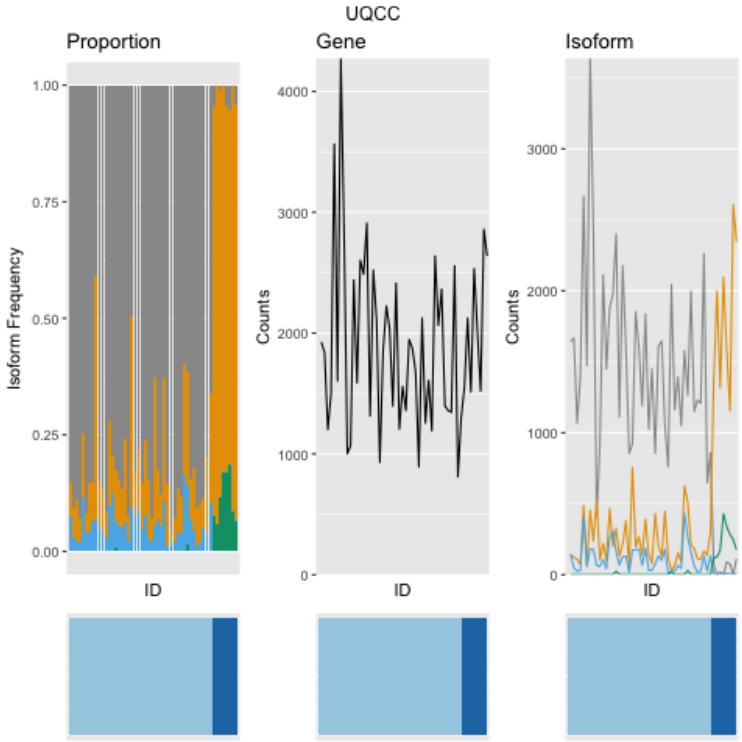


Figure 5.12: **Relative isoform frequency, gene and isoform levels of UQCC** The dark blue group represents a set of UVM patients with mutations in *SF3B1* who were shown to cluster together using proportion clustering. In the left most figure, we plotted the relative isoform frequency of each isoform. The x-axis is each individual and each isoform is represented by a different color. The proportion of different colors in the columns denotes the relative frequency of each isoform. The middle figure shows gene expression, and the x-axis is each individual. The right most figure shows isoform expression. Again, the x-axis is each individual, and each isoform is represented by a different color. We note the isoform that is preferentially expressed switches dramatically, though this does not manifest as a noticeable change in gene expression. Previous work by (Furney et al., 2013) found UQCC to be differentially expressed in the presence of a *SF3B1* mutation. We also found several isoforms of this gene to be differentially expressed with respect to the *SF3B1* mutation.

Overlap Groups	Groups P1	P2	P3
P1	–	0.077	0.064
P2	–	–	0.059
P3	–	–	–

(a) Proportion Clustering

Overlap Groups	Groups I1	I2	I3
I1	–	0.018	0.04
I2	–	–	0.036
I3	–	–	–

(b) Isoform Clustering

Table 5.2: **Comparison at $K = 3$ in the TCGA LAML data set:** The proportion of genes sharing contrasting isoforms between the clusters in proportion clustering is higher than between the clusters in isoform clustering

including *TP53* and *RUNX1*. In the TCGA UVM data set, we found a very strong signal due to differential isoform expression at $K = 2$. Within each of these resulting clusters, we showed that individuals with *SF3B1* mutations cluster strongly together.

Overlap Groups	Groups P1	P2	P3	P4
P1	–	0.073	0.062	0.033
P2	–	–	0.069	0.089
P3	–	–	–	0.095
P4	–	–	–	–

(a) Proportion Clustering

Overlap Groups	Groups I1	I2	I3	I4
I1	–	0.015	0.041	0.026
I2	–	–	0.037	0.024
I3	–	–	–	0.027
I4	–	–	–	–

(b) Isoform Clustering

Table 5.3: **Comparison at $K = 4$ in the TCGA LAML data set:** The proportion of genes sharing contrasting isoforms between the clusters in proportion clustering is higher than between the clusters in isoform clustering

Overlap Groups	Groups P1	P2	P3	P4	P5
P1	–	0.021	0.018	0.054	0.050
P2	–	–	0.037	0.095	0.092
P3	–	–	–	0.063	0.054
P4	–	–	–	–	0.065
P5	–	–	–	–	–

(a) Proportion Clustering

Overlap Groups	Groups I1	I2	I3	I4	I5
I1	–	0.014	0.020	0.013	0.022
I2	–	–	0.051	0.013	0.029
I3	–	–	–	0.028	0.027
I4	–	–	–	–	0.018
I5	–	–	–		

(b) Isoform Clustering

Table 5.4: **Comparison at $K = 5$ in the TCGA LAML data set:** The proportion of genes sharing contrasting isoforms between the clusters in proportion clustering is higher than between the clusters in isoform clustering

Bibliography

Aggarwal, C C and C K Reddy (2013). *Data Clustering: Algorithms and Applications*. 1st. Chapman & Hall/CRC. ISBN: 1466558210, 9781466558212.

Alizadeh, A A et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." In: *Nature* 403.6769, pp. 503–511.

Anders, S (2010). "Analysing RNA-Seq data with the "DESeq" package". In: *Molecular Biology* 43.4, pp. 1–17.

Anders, S, A Reyes, and W Huber (2012). "Detecting differential usage of exons from RNA-seq data." In: *Genome research* 22.10, pp. 2008–2017.

Anderson, M J (2001). "A new method for non-parametric multivariate analysis of variance". In: *Austral Ecology* 26.1, pp. 32–46.

Berninger, P et al. (2008). "Computational analysis of small RNA cloning data." In: *Methods (San Diego, Calif.)* 44.1, pp. 13–21.

Brunet, J-P et al. (2004). "Metagenes and molecular pattern discovery using matrix factorization." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.12, pp. 4164–4169.

Bu, J, X Chi, and Z Jin (2013). "HSA: a heuristic splice alignment tool." In: *BMC systems biology* 7 Suppl 2, S10.

Bullard, J H et al. (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." In: *BMC bioinformatics* 11, p. 94.

Burset, M, I A Seledtsov, and V V Solovyev (2000). "Analysis of canonical and non-canonical splice sites in mammalian genomes." In: *Nucleic Acids Research* 28.21, pp. 4364–4375.

Cancer Genome Atlas Research Network (2012). "Comprehensive molecular portraits of human breast tumours." In: *Nature* 490.7418, pp. 61–70.

— (2013). "Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia". In: *The New England journal of medicine* 368.22, pp. 2059–2074.

Cancer Genome Atlas Research Network (2014). "Comprehensive molecular profiling of lung adenocarcinoma." In: *Nature* 511.7511, pp. 543–550.

Chen, M and J L Manley (2009). "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." In: *Nature reviews. Molecular cell biology* 10.11, pp. 741–754.

DeBoever, C et al. (2015). "Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers." In: *PLoS computational biology* 11.3, e1004105.

Denoeud, F et al. (2008). "Annotating genomes with massive-scale RNA sequencing". In: *Genome biology* 9.12, R175.

Deza, M M and E Deza (2014). *Encyclopedia of Distances*.

Efron, Bradley (1986). "Double Exponential Families and Their Use in Generalized Linear Regression". In: *Journal of the American Statistical Association* 81.395, pp. 709–721.

Everitt, B S et al. (2011). *Cluster Analysis*.

Ferreira, P G et al. (2014). "Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia". In: *Genome Research* 24.2, pp. 212–226.

Filippone, M et al. (2008). "A survey of kernel and spectral methods for clustering - ScienceDirect". In: *Pattern recognition*.

Fu, X-D and M Ares Jr (2014). "Context-dependent control of alternative splicing by RNA-binding proteins". In: *Nature Reviews Genetics* 15.10, pp. 689–701.

Furney, S J et al. (2013). "SF3B1 mutations are associated with alternative splicing in uveal melanoma." In: *Cancer discovery* 3.10, pp. 1122–1129.

Gao, K et al. (2008). "Human branch point consensus sequence is yUnAy". In: *Nucleic Acids Research* 36.7, pp. 2257–2267.

Gentien, D et al. (2014). "A common alternative splicing signature is associated with SF3B1 mutations in malignancies from different cell lineages." In: *Leukemia* 28, pp. 1355–1357.

Gönen, M and E Alpaydın (2011). "Multiple Kernel Learning Algorithms". In: *Journal of Machine Learning Research* 12.Jul, pp. 2211–2268.

Gonzàlez-Porta, M et al. (2012). "Estimation of alternative splicing variability in human populations." In: *Genome research* 22.3, pp. 528–538.

Hammerman, P S et al. (2012). "Comprehensive genomic characterization of squamous cell lung cancers". In: *Nature* 489.7417, pp. 519–525.

Hartley, S W and J C Mullikin (2016). "Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq." In: *Nucleic acids research*.

Huang, G T et al. (2012). "Spectral clustering strategies for heterogeneous disease expression data". In: *Proceedings of the Pacific Symposium*. World Scientific, pp. 212–223.

Ilagan, Janine O et al. (2015). "U2AF1 mutations alter splice site recognition in hematological malignancies." In: *Genome research* 25.1, pp. 14–26.

Jiang, H and W H Wong (2009). "Statistical inferences for isoform expression in RNA-Seq". In: *Bioinformatics* 25.8, pp. 1026–1032.

Johnson, W E, C Li, and A Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." In: *Biostatistics (Oxford, England)* 8.1, pp. 118–127.

Katz, Y et al. (2010). "Analysis and design of RNA sequencing experiments for identifying isoform regulation." In: *Nature Methods* 7.12, pp. 1009–1015.

Kim, Jung-Hyun et al. (2016). "SON and Its Alternatively Spliced Isoforms Control MLL Complex-Mediated H3K4me3 and Transcription of Leukemia-Associated Genes." In: *Molecular cell* 61.6, pp. 859–873.

Lavi, Ofer, Gideon Dror, and Ron Shamir (2012). "Network-induced classification kernels for gene expression profile analysis." In: *J Comput Biol* 19.6, pp. 694–709.

Leek, J T et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." In: *Nature reviews. Genetics* 11.10, pp. 733–739.

Leng, N et al. (2013). "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments." In: *Bioinformatics* 29.8, pp. 1035–1043.

Lewandowska, M A (2013). "The missing puzzle piece: splicing mutations." In: *International journal of clinical and experimental pathology* 6.12, pp. 2675–2682.

Li, B and C N Dewey (2011). "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC Bioinformatics* 12.1, p. 323.

Li, W and T Jiang (2012). "Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads". In: *Bioinformatics (Oxford, England)* 28.22, pp. 2914–2921.

Li, Y et al. (2015). "RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq." In: *PLoS ONE* 10.9, e0136653.

Luxburg, U von (2007). "A tutorial on spectral clustering". In: *Statistics and Computing* 17.4, pp. 395–416.

Marioni, J C et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." In: *Genome research* 18.9, pp. 1509–1517.

Matera, A G and Z Wang (2014). "A day in the life of the spliceosome." In: *Nature reviews. Molecular cell biology* 15.2, pp. 108–121.

McNicholas, P D and T B Murphy (2010). "Model-based clustering of microarray expression data via latent Gaussian mixture models." In: *Bioinformatics (Oxford, England)* 26.21, pp. 2705–2712.

Monlong, J et al. (2014). "Identification of genetic variants associated with alternative splicing using sQTLseekeR." In: *Nature communications* 5, p. 4698.

Monti, S et al. (2003). "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data". In: *Machine learning* 52.1-2, pp. 91–118.

Mortazavi, A et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7, pp. 621–628.

Ng, A Y, M I Jordan, and Y Weiss (2001). "On Spectral Clustering: Analysis and an algorithm". In: *Advances in Neural Information Processing Systems*. MIT Press, pp. 849–856.

Nilsen, T W and B R Graveley (2010). "Expansion of the eukaryotic proteome by alternative splicing." In: *Nature* 463.7280, pp. 457–463.

Nowak, D G et al. (2008). "Expression of pro- and anti-angiogenic isoforms of VEGF is differentially regulated by splicing and growth factors." In: *Journal of cell science* 121.Pt 20, pp. 3487–3495.

Oltean, S and D O Bates (2014). "Hallmarks of alternative splicing in cancer". In: *Oncogene* 33.46, pp. 5311–5318.

Pan, Q et al. (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing". In: *Nat Genet* 40.12, pp. 1413–1415.

Perou, C M et al. (2000). "Molecular portraits of human breast tumours." In: *Nature* 406.6797, pp. 747–752.

Przychodzen, Bartlomiej et al. (2013). "Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms." In: *Blood* 122.6, pp. 999–1006.

Qiu, Y et al. (2009). "The anti-angiogenic isoforms of VEGF in health and disease." In: *Biochemical Society transactions* 37.Pt 6, pp. 1207–1213.

Qu, Y and S Xu (2004). "Supervised cluster analysis for microarray data based on multivariate Gaussian mixture." In: *Bioinformatics (Oxford, England)* 20.12, pp. 1905–1913.

Quesada, V et al. (2012). "Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia." In: *Nature Genetics* 44.1, pp. 47–52.

Richard, H et al. (2010). "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments". In: *Nucleic Acids Research* 38.10, e112.

Roberts, A et al. (2011). "Improving RNA-Seq expression estimates by correcting for fragment bias". In: *Genome Biology* 12.3, R22.

Robinson, M D, D J McCarthy, and G K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics (Oxford, England)* 26.1, pp. 139–140.

Ruddy, S, M Johnson, and E Purdom (2016). "Shrinkage of dispersion parameters in the binomial family, with application to differential exon skipping". In: *The Annals of Applied Statistics* 10.2, pp. 690–725.

Salzman, J, H Jiang, and W H Wong (2010). *Statistical Modeling of RNA-Seq Data*. Tech. rep. Palo Alto.

Shen, S et al. (2016). "SURVIV for survival analysis of mRNA isoform variation." In: *Nature Communications* 7, p. 11548.

Shi, J and J Malik (2000). "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8, pp. 888–905.

Si, Y et al. (2014). "Model-based clustering for RNA-seq data." In: *Bioinformatics (Oxford, England)* 30.2, pp. 197–205.

Sigurgeirsson, B, O Emanuelsson, and J Lundeberg (2014). "Sequencing Degraded RNA Addressed by 3' Tag Counting". In: *PLOS ONE* 9.3, e91851.

Singh, R K and T A Cooper (2012). "Pre-mRNA splicing in disease and therapeutics". In: *Trends in molecular medicine* 18.8, pp. 472–482.

Song, Q, S D Merajver, and J Z Li (2015). "Cancer classification in the genomic era: five contemporary problems." In: *Human genomics* 9.1, p. 27.

Sorlie, T et al. (2001). "Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications". In: *Proceedings of the National Academy of Sciences of the United States of America* 98.19, pp. 10869–10874.

Souto, M CP de et al. (2008). "Clustering cancer gene expression data: a comparative study". In: *BMC Bioinformatics* 9.1, p. 1.

Sveen, A et al. (2016). "Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes." In: *Oncogene* 35.19, pp. 2413–2427.

Tamayo, P et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." In: *Proceedings of the National Academy of Sciences of the United States of America* 96.6, pp. 2907–2912.

Tarca, A L, R Romero, and S Draghici (2006). "Analysis of microarray experiments of gene expression profiling." In: *American journal of obstetrics and gynecology* 195.2, pp. 373–388.

Trapnell, C, D G Hendrickson, et al. (2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nature Biotechnology* 31.1, pp. 46–53.

Trapnell, C, B A Williams, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28.5, p. 511.

Venables, J P (2004). "Aberrant and Alternative Splicing in Cancer". In: *Cancer Research* 64.21, pp. 7647–7654.

Wang, B et al. (2016). "Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing." In: *Nature communications* 7, p. 11708.

Wang, E T et al. (2008). "Alternative isoform regulation in human tissue transcriptomes". In: *Nature* 456.7221, pp. 470–476.

Wang, K et al. (2010). "MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery". In: *Nucleic Acids Research* 38.18, e178–e178.

Wang, L, S Wang, and W Li (2012). "RSeQC: quality control of RNA-seq experiments". In: *Bioinformatics (Oxford, England)* 28.16, pp. 2184–2185.

Witten, D M (2011). "Classification and clustering of sequencing data using a Poisson model". In: *The Annals of Applied Statistics* 5.4, pp. 2493–2518.

Witten, D M and R Tibshirani (2010). "A framework for feature selection in clustering." In: *Journal of the American Statistical Association* 105.490, pp. 713–726.

Wu, Z, X Wang, and X Zhang (2011). "Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq". In: *Bioinformatics* 27.4, pp. 502–508.

Yokoyama, Akihiko et al. (2005). "The menin tumor suppressor protein is an essential oncogenic cofactor for MLL-associated leukemogenesis." In: *Cell* 123.2, pp. 207–218.

Zeng, H and Y Cheung (2011). "Feature selection and kernel learning for local learning-based clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8, pp. 1532–1547.