Investigations into the contribution of retrotransposon activation in neurodegenerative disease

Kathryn Shea O'Neill

Advisor: Dr. Molly Gale Hammell

Committee Members: Drs. Adam Siepel, Camila Dos Santos, and Rob Martienssen

External Examiner: Dr. Hemali Phatnani

Cold Spring Harbor School of Biological Sciences

Table of Contents

List of Tables and Figures

List of Abbreviations

AE – autoencoder

ALS – Amyotrophic Lateral Sclerosis

EM – expectation maximization

ERV – endogenous retrovirus

FTD - Frontotemporal Dementia

FTD-TDP – Frontotemporal Dementia with TDP-43 proteinopathy

GMM – Gaussian mixture model

LINE – long interspersed nuclear element

LTR – long terminal repeat

ML – machine learning

NMF – non-negative matrix factorization

NN – neural net

PCA – principal components analysis

RTE – retrotranspososable element

scRNA-seq – single cell RNA-seq

SINE – short interspersed nuclear element

TE – transposable element

WGCNA – weighted gene co-expression network analysis

Dedication and Acknowledgements

I Kathryn Shea O'Neill, daughter of Roshaun Felice Fisher and Kevin Kirk O'Neill, dedicate the culminative labor of this work, merits and flaws, to the incredible support I have received throughout my life and tenure as a graduate student. I deeply thank my mother for bestowing to me her ability to burn what is needed to move forward and protect life, my father for his serenity in the face of death, and both for their honest love and humanity. I thank Dr. Molly Gale Hammell for her profound dedication to mentorship, incisive intellect, unwavering belief in me, and graceful composure through my development as a young scientist. Her role in the success of this dissertation cannot be overstated, and her ability to cultivate so much compassion in her work and workplace is an inspiration. I thank my committee members, Drs. Camila Dos Santos, Adam Siepel, and Rob Martienssen, for their insightful comments and mentorship. I thank Peter Koo for taking the time to explain how to build a good neural network. I thank John Andrews for his reverence for the intensity of life, and for his heart which can hold such heat and remain tender. I thank John's parents for their resilient kindness in the face of incredible odds. I thank Dr. Kenya Mitchell and the UC Davis McNair Scholars program for catalyzing my pursuit of a graduate degree, and for guiding me to step into my true potential. I thank all of my friends for their love and predilection for experiential rollercoasters, sometimes even actual rollercoasters. I thank the Cold Spring Harbor Laboratory School of Biological Sciences and all of its staff for their support and the deep teachings they offered me about the academy. I thank Roberta Swanson, Julia Herrera, and Julio Eleamar Barrios Durán for opening my eyes to the global kaleidoscope of humanity. I thank Lincoyán Francisco Berríos for showing me the first step on my path towards decolonization, and to him and his daughter Ayelén for encouraging me to continue to create, in spite all doubt, because it is necessary. I thank the land, and its original custodians the Matinecock people, upon which this work was conducted for graciously supporting us and the structures we rely upon to perform our works. Finally, I dedicate this thesis, and my infatuation with biology to evolution and the power of the will to live. I am humbly liberated by the unfathomable strength latent in and exemplified by all living organisms, without exception. I hope that rapture for the diversity of responses to conflict in our world, and the oceans of death that have structured the foundation from which we act lead us to recognize our innate unity. May interdependence forever keep us resolute in our struggle to find home and wholeness.

Chapter 1: Introduction

1.1 Transposons and their role in disease

Transposable elements (TEs) are mobile genomic elements that have (or once had) the ability to replicate themselves within a host genome. TEs have a rich history since their discovery by Barbara McClintock in *Zea mays*, and continue to be source of great interest to scientists. They are an incredibly diverse group of genetic elements that have co-evolved with their host genomes. Hosts have responded with increasingly intricate mechanisms by which to regulate transposons, and concomitantly gene expression activity. This elaborate competition between the drive of TEs to replicate, and for the host genome to protect itself, is a source of intraorganismal competition that over time has resulted in our current genetic regulatory landscapes (Wang et al. 2007; Feschotte 2008; Bourque et al. 2008; Schmidt et al. 2012; Trizzino et al. 2017). Consequently, transposons comprise a significant proportion of the genome, at least 45% in humans alone (Lander et al. 2001). Importantly, TEs carry with them cis-regulatory elements to promote their own expression, and in many cases these cis-regulatory elements have been exapted by the host genome to serve as regulatory elements for host genes. One example that is particularly relevant for human health, involves the development of cis-regulatory elements responsive to interferon-gamma pathways that regulate genes involved in mammalian innate immunity. In primates, regulatory elements from the MER41 endogenous retroviruses have been coopted to serve as gene enhancers bound by STAT1 and IRF1, two transcription factors critical for gene expression activation by interferon in humans (Chuong et al. 2016). Thus, an endogenous retrovirus that previously invaded our primate ancestors has been repurposed to guard against new threats. Examples such as this give a brief window into the complicated roles that TEs play in human health and disease, and the evolutionary importance of TEs in the shaping gene expression networks. However, not all TE activity is beneficial or benign.

TEs first and foremost are genetic parasites and mutagens, which is why the genome has developed strong defenses against them to be elaborated on later in this introduction. Briefly, a transposon can create new genomic insertions through one of two mechanisms, either through a "cut and paste" mechanism or a "copy and paste" mechanism. The "cut and paste" mechanism is used by Class II DNA transposons. These transposons directly excise themselves from the genome and integrate at new locations using an encoded transposase protein. Class II transposons only comprise around 2.8% of the TEs in the human genome, are no longer transpositionally active in the human genome, and are thus relatively ancient compared to Class I transposons (Cordaux and Batzer 2009).The rest of this thesis will focus on the larger group of TEs known as retrotransposons, or Class I elements.

Class I retrotransposons (RTEs) used a "copy and paste" mechanism that first produces an RNA transcript of the RTE; this transcript is then reverse transcribed into cDNA and later integrated into the host genome, leaving the original DNA copy of the RTE intact. There is significant variety in the mechanisms by which different types of retrotransposons achieve transposition, but the elements can be divided into three main categories: long terminal repeat (LTR) retrotransposons, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). LTR retrotransposons include the previously discussed endogenous retroviruses (ERVs), such as MER41, and typically encode three main proteins: *gag pol* and *env*. ERVs and, LTR retrotransposons more generally, share a rich and complex evolutionary history with exogenous retroviruses (Johnson 2019), though only ERVs are the result of ancient viral infections of the germline. LINEs encode two proteins ORF1, an RNA binding protein, and ORF2 a reverse transcriptase and integrase that allows for cDNA production and ultimately integration. The ORF2 protein from these LINE elements, has the potential to transpose the otherwise non-autonomous third major class of RTEs, short interspersed nuclear elements (SINEs). The most common transposon in the human genome, is a SINE/Alu element, present at over a million copies.

Most human RTEs are currently inactive, in terms of their ability to transpose. The only TE known to autonomously transpose itself is a subfamily of LINE elements: L1HS (LINE-1 Human Specific). However, the ORF2 protein produced by actively transcribed and translated L1HS elements is available to allow for transpositional activity of other non-autonomous elements, particularly Alu elements, of which the most active subfamilies include AluYa5 and AluSx (Bennett et al. 2008). This transposition activity was largely thought to be limited to the germline where transposons are most highly expressed. More recently, somatic transposition events have been observed in long lived post-mitotic cells like neurons, where insertional events are reported to occur at a rate between 0.04 and 13 insertions per cell (Tam et al. 2019a). While it is currently unknown whether most somatic and germline transposition events will impact cellular function, 126 examples of novel RTE insertion events leading to human disease have been catalogued (Hancks and Kazazian 2016). Broadly, each of these insertions caused disease through insertion into a protein-coding gene locus, which led to either an exonic coding mutation, improper splicing of the host gene, or mislocalization of the encoded protein. RTE transposition is not the only way for a transposon impact cellular function. Next, I will describe alternate ways in which TEs can disturb the cell and contribute to human disease.

As mentioned earlier, TEs have played a critical role in the development of the regulatory structure of the human genome. Therefore, the genome must strike a delicate balance between expression and repression TEs. If a cell becomes overly permissive of TE expression, their promoters can lead to aberrant transcriptional activity of adjacent host genes. This becomes increasingly important in cellular contexts where TE expression is derepressed, as in senescent cells or some cancers (Bourque et al. 2018). The link between TE derepression and senescence is of particular interest in the context of neurodegenerative disease, where long-lived post-mitotic cells have been described to enter senescent-like states (Dehkordi et al. 2021).

Additionally, TE transcriptional and translational byproducts can activate innate immune signaling pathways. The mammalian cell has multiple nucleic acid sensors which can trigger downstream proinflammatory signaling cascades. Briefly, the cGAS/STING pathway can be activated by cytosolic cDNA produced via TE encoded reverse transcriptases, such as LINE-1 ORF2. RIG-I like receptors (RLRs) which have affinity for short dsRNA and ssRNA and have been shown in vitro to bind Alu elements. TE derived ssRNAs and dsRNAs can also activate Toll-like receptors, including TLR3. Examples of this TE-induced innate immune activation have been observed in the contexts of cancer, senescence, and Aicardi-Goutières syndrome (AGS). AGS is a particularly interesting example of these phenomena for this thesis as AGS specifically affects the brain, spinal cord and immune system. Finally, some HERV encoded proteins have been shown to be cytotoxic when highly expressed, with Env proteins from HERV-W and HERV-K ERVs implicated in multiple sclerosis and amyotrophic lateral sclerosis, respectively (Li et al. 2015; Bourque et al. 2018; Gázquez-Gutiérrez et al. 2021). Through these mechanisms, we see that TE biology is inextricably linked to the biology of the cell, and affords many mechanisms by which TEs may potentiate disease. This contentious relationship has spurred the development of multiple mechanisms by which the cell seeks to control TE expression.

Mammalian TE expression is jointly regulated at both the transcriptional and post-transcriptional levels. Transcriptionally, TEs are controlled by DNA methylation, histone modifications, and repressive transcription factors, such as the Krüppel-associated box domain zinc finger (KRAB-ZNF) proteins (Xie et al. 2013; Friedli and Trono 2015). Post-transcriptional control is achieved by several small RNA mediated interference pathways: Piwi-interacting RNAs (piRNAs) mediate TE silencing in the germline, small interfering RNAs (siRNAs) contribute to suppression in both the germline and somatic cells, and 3' tRNA fragments (3' tRFs) suppress the activty of ERVs. In the mammalian germline, transcriptional and post-transcriptional silencing pathways are coupled through the activity of the piRNA pathway, which uses piRNAs to guide DNA methylation complexes to TE genomic loci. This piRNA-coupled DNA methylation mechanism plays a critical role in TE silencing during early embryogenesis when genomic DNA methylation is reset (Aravin et al. 2008; Miyoshi et al. 2016). I will briefly discuss piRNAs when describing the TEsmall pipeline later in this work, as piRNAs are critical to TE silencing in the germline. However, the bulk of this work is limited to research on somatic cells and will focus on the mechanisms used to silence TEs in those contexts.

In somatic cells, TE silencing is largely mediated through transcriptional pathways, including DNA methylation and repressive histone environments (Xie et al. 2013; Friedli and Trono 2015). Transcriptional control is established early during embryogenesis and cellular differentiation by de novo DNA methyltransferases through the formation of 5-methylcytosine by DNMT3A and DNMT3B. DNA methylation patterns are then maintained over the lifetime of the cell through maintenance methyltransaferases like DNMT1 (Moore et al. 2013; Edwards et al. 2017). These DNA methylation marks additionally interact with other protein complexes to facilitate transcriptional repression via repressive histone marks, such as H3K9me3 and H3K27me3 (Hyun et al. 2017). These histone modifications are used to mark TE genomic regions through several chromatin remodeling pathways, including SETDB1, PRC2, and the HUSH complex (Hyun et al. 2017; Allshire and Madhani 2018; Déléris et al. 2021; Seczynska et al. 2022). These histone-based pathways can be more dynamic and responsive to cellular state as the genome opens and closes around genic and gene-regulatory regions to enable gene expression changes that respond to cellular stimuli (Hyun et al. 2017; Allshire and Madhani 2018).

Somatic cells also use post-transcriptional pathways to ensure multiple redundant levels of TE control. For example, the L1HS LINE elements encode bidirectional promoters that can create long dsRNAs, and these dsRNAs in turn become substrates for generation of siRNAs that target L1HS transcripts (Svoboda et al. 2004; Soifer et al. 2005). Briefly, siRNA silencing is achieved by a very similar mechanism to the miRNA processing pathway. Cytoplasmic dsRNAs are loaded into Dicer that cleave the dsRNA into short ~21nt small RNA duplexes (the siRNAs). These siRNAs are then loaded into an Argonaute protein, most commonly AGO2 in humans. One RNA strand of the siRNA duplex is degraded, while the other is used to target complementary mRNAs as part of an RNA-Induced Silencing Complex (RISC). This can lead to slicing of the target RNA by AGO2 containing RISC proteins. Alternately, RISC complexes may silence their targets through translational repression, mRNA decapping and deadenylation, and ultimately degradation of the RNA target by exonucleases (Wilson and Doudna 2013).

The final post-transcriptional mechanism by which TEs are repressed involves the activity of 3′ tRNA fragements (3′ tRFs), originally described by Schorn et al. (Schorn et al. 2017). This mechanism is specific to ERVs, which typically use host tRNA sequences to prime their reverse transcription reactions. These 3′ tRFs bind to sequences within ERV transcripts called the primer binding site (PBS) and block access by full-length tRNAs to those PBS sequences, thus blocking the ability of RT enzymes to produce cDNAs from ERV RNA transcripts. While these 3′ tRFs do act post-transcriptionally to silence ERV activity, they do not act to regulate the level of ERV RNA transcripts, as siRNAs and piRNAs do on other TE transcripts. Thus, 3′ tRFs are predominantly effective against ERVs that retain the ability to produce cDNA, providing an additional level of control for the youngest ERV elements.

TE repression is a rich field of study, and is inextricably linked to how gene expression is regulated generally. This thesis is dedicated in part to understanding the transcriptional landscape of the brain in amyotrophic lateral sclerosis (ALS), and deregulation of RTEs in that context. Therefore, it is important to understand the basic rules by which the mammalian somatic cell would normally address RTE expression, to contextualize that landscape.

1.2 TDP-43 loss of function leads to retrotransposon de-silencing

TAR DNA-binding protein 43 (TDP-43) is a ubiquitously expressed protein associated capable of binding DNA, RNA, and other proteins. TDP-43 has been linked to multiple neurodegenerative diseases, forming aggregate proteinopathies in the brain in over 90% of ALS cases and up to 50% of frontotemporal dementia (FTD) cases (Jo et al. 2020). In order to understand the role TDP-43 aggregates play in these diseases, researchers have been interested in determining how these proteinopathies affect TDP-43's normal function, and whether or not these aggregates result in a loss of TDP-43 function, a novel toxic gain of function, or a combination of these effects. TDP-43 is normally located in both the cytoplasm and the nucleus, but when functioning normally spends most of its time in the nuclear compartment, unphosphorylated. TDP-43 proteinopathies occur when the majority of this protein is present in a hyper-phosphorylated and aggregate form in the cytoplasm, suggesting that aggregation results in both a nuclear loss of function as well as a toxic gain of function in the cytoplasm (Vanden Broeck et al. 2014). To substantiate these models for TDP-43 proteinopathy, several studies have perturbed TDP-43 activity through either overexpression and/or knockdown of TDP-43. Overexpression of human TDP-43 in *D. melanogaster* glial and neuronal cells showed increased TDP-43 expression in fly brains leads to cytoplasmic mislocalization of both the over-expressed hTDP-43 as well as the endogenous fly ortholog, TBPH (Krug et al. 2017). This dominant-negative loss of TDP-43 function in the fly brain led to degeneration as well as de-silencing of RTE expression. Additionally, this publication showed that expression of hTDP-43 caused a reduction in the capacity for Dicer-2/Ago2 siRNA mediated mRNA silencing in glial and mushroom body neurons in *Drosophila*. Early data from our lab using short hairpin knockdown assays targeting TDP-43 in human somatic cells showed RTEs are globally upregulated upon TDP-43 loss of function, and this data is presented later in Chapter 3. These data taken in concert show that loss of normal, nuclear TDP-43 function can lead to RTE activation, potentially through a reduction in the capacity for siRNA mediated silencing of RTEs. The *D. melanogaster* study did assess the cellular localization of TDP-43, and showed the overexpression of hTDP-43 caused nuclear clearance of hTDP-43, and cytoplasmic accumulation of phosphorylated hTDP-43 consistent with pathological presentation. While this presentation, does not separate the conditions of a loss of function or gain of function mutation as there is both nuclear clearance and cytoplasmic aggregation, this presentation is consistent with the presentation of TDP-43 pathology in human ALS. This clinical relevance further supports the investigation of siRNA biology as it relates to TDP-43 biology and ALS.

Separating the gain of function and loss of function characteristic of TDP-43 is still an area of active research TDP-43 is a highly pleiotropic gene, however, as a loss of function perturbation is sufficient to upregulate RTE expression in human somatic cells we will limit the focus of our molecular assays to RTE activation by TDP-43 loss of function. In regards to analysis of patient derived samples with TDP-43 proteinopathy, discussion of TDP-43 associated dysfunction must be held from the perspective that it is still unclear how these aggregates relate to normal TDP-43 function and the general etiology of neurodegenerative disease (Suk and Rousseaux 2020).

1.3 Description of ALS and FTD-TDP

ALS and FTD-TDP are related diseases via their shared TDP-43 proteinopathy. As TDP-43 dysfunction is known to be associated with RTE derepression, it is of interest to investigate RTE activity in the context of these diseases. In this section, I will describe both of these diseases to contextualize Chapters 4 and 5 of this work.

ALS is a terminal neurodegenerative disease which primarily targets motor neurons, and currently has no cure. Symptoms of this disease include loss of motor function, most devastatingly loss of speech, ability to swallow, and breathe. Death predominantly results from loss of respiratory function, or from complications associated with decreased respiratory function like pneumonia. The average age of onset for ALS is 50 years of age (Keon et al. 2021), and is a fast-progressing disease, with median times from ALS diagnosis to death of 2-4 years. However, some patients live much longer with the disease (Hardiman et al. 2017). Taken together, ALS is a devastating disorder which is still poorly understood for a variety of reasons.

ALS is a largely sporadic (sALS) disease with approximately 80-90% of cases having no associated genetic mutation or family history of the disorder. In the small fraction of familial cases (~10%), the known genetic drivers of ALS are often incompletely penetrant (Hardiman et al. 2017). In recent years, this has led researchers toward the hypothesis that ALS functions more as a destination reached through a combination of genetic and environmental drivers within the nervous system (Keon et al. 2021). These factors tend to converge on axonal degeneration in motor neurons, but evidence suggests that there are many cell types involved in this process, and we cannot limit our investigation of ALS to neurons (Keon et al. 2021). The most common gene perturbations associated with ALS are repeat expansions within *C9orf72*, and mutations in *SOD1*, *FUS*, and *TARDPB* (the gene encoding TDP-43) (Hardiman et al. 2017). However, many genes have been associated with ALS, and can be grouped into several categories based on cellular function (Hardiman et al. 2017; Ghasemi and Brown Jr. 2018). These functions include: protein stability and degradation, RNA metabolism, chromatin biology, and axonal and cytoskeletal biology. Additionally, mutations in genes associated with these themes, according to Ghasemi and Brown Jr. (Ghasemi and Brown Jr. 2018), are hypothesized to cause downstream secondary effects in other cellular processes like ER stress, autophagy, mitochondrial function, axonal transport, proteosomal function, neuronal excitotoxicity, and neuroinflammation. Notably, ALS associated genes touch nearly every arm of cell biology resulting in highly complex genetic interactions and changes in metabolic state. This makes ALS a very difficult disease to study, and currently requires better characterization of how these varied genetic and metabolic perturbations converge on the disease known as ALS.

As noted above, ALS also shares genetic and phenotypic overlap with a related disorder, fronto-temporal dementia with TDP-43 inclusions (FTD-TDP). FTD is a neurodegenerative disease that affects cortical neurons of the frontal and temporal lobes. FTD is an earlier onset neurodegenerative disease, with a median age at diagnosis between 45-65. In contrast to ALS patients, however, patients with FTD often live up to 20 years post diagnosis. FTD patients can be broadly separated into those with evidence of Tau aggregates in cortical tissues (FTD-Tau) and those that predominantly show aggregates of TDP-43 (FTD-TDP). As outlined by Ghasemi and Brown Jr., it is the FTD-TDP subset that shares many causal genetic factors with ALS. Both diseases can be caused by mutations in C9orf72, TARDBP, VCP, SQSTMI, UBQLN2, and CHMP2B (Ghasemi and Brown Jr. 2018). Moreover, many ALS patients will eventually develop cognitive and behavioral symptoms consistent with FTD, and many FTD-TDP patients also develop motor deficits. This has led researchers to unite the two diseases under the common umbrella term "ALS/FTD Spectrum Disorder." This invites questions about how interactions between cell type specific vulnerabilities, environmental factors, and genetic background, converge upon these two syndromes. This is also an emerging theme in neurodegenerative diseases more broadly, as proteinopathies associated with multiple neurodegenerative diseases are often found to be concomitant (Robinson et al. 2018).

To address the heterogeneity associated with ALS and FTD-TDP, my lab previously sought to investigate if there were molecular subtypes of ALS that might allow us to better characterize this disease. They found that there are three dominant molecular subtypes present in the transcriptomes of post mortem cortical tissue collected from ALS patients. These three groups include an oxidative stress group, a microglial activation group, and a TDP-43 pathology associated group (Tam et al. 2019b). However, this was a pilot study conducted on 148 samples from 77 patients using non-negative matrix factorization as a clustering method to separate distinct subtypes. Building new statistical models to expand upon this study and characterize ALS subtypes will be the focus of Chapter 4 of this thesis. The next section will introduce the types of algorithms which were previously used to identify ALS subtypes and discuss their limitations.

1.4 Introduction to relevant machine learning techniques to detect structure within biological sequencing data

1.4.1 – Broad classes of machine learning methods & applications

Machine learning is an umbrella term for a class of statistical methods which rely most heavily on data to construct a statistical model, rather than on a scientist's understanding of the data to create an appropriate model. This 'data first' perspective is a powerful approach to detect, predict, and/or interpret factors of interest within a given dataset. This paradigm makes machine learning an appealing approach to apply to large datasets generated from complex and difficult to model systems, such as those seen in biology.

Broadly, machine learning methods can be divided into two categories, supervised and unsupervised. Both types of methods can be powerful tools for biological discovery, and both will be applied to the projects described in this dissertation. Supervised learning is a type of task which requires known labels for the input data and/or for the output variables of interest. Classifiers are a group of supervised learning methods that assign data to specific categories, and include well-known methods such as: neural nets, decision trees, and support vector machines. Alternately, regression models that learn relationships between input and output variables (e.g., linear and logistic regression) also fall into the class of supervised learning methods. Unsupervised learning methods describe algorithms that are designed to learn structure shared across data points in order to better understand how aspects of the data are related. Clustering algorithms are probably the best known of the unsupervised learning methods, but dimensionality reduction and recommendation engines also fit into this group.

An example task for a supervised learning method might use regression to measure the relationship between outdoor temperature and household energy consumption, where the daily temperature represents the input variable, and energy consumption of a household the output variable. A person interested in learning to predict this relationship would record both factors for a particular household over a period of time, and based upon this data, predict the household's energy consumption for a given future temperature outside. One can imagine more complicated examples which take multiple inputs (i.e. income, household square footage, or time of year) to predict energy consumption more generally. These more complicated models might determine the weights of each contributing factor to overall energy consumption and how to combine these factors into a single prediction. This temperature/energy example may also be reimagined as a classification problem with a categorical output variable. Instead of predicting energy consumption as a quantitative variable, one might be interested in predicting what class of clothing the homeowner was going to wear (shorts or pants) based on temperature.

In each of these toy examples for supervised learning methods, a large set of labeled input and output data (temperature and energy/clothing) would have to be collected previously to train the model, hence the name *supervised* learning. The models discussed in Chapter 4 to classify ALS patients into molecular subtypes are examples of supervised models. The specific supervised method used in this thesis will be described in greater detail in section 1.4.8.

As described briefly above, unsupervised learning methods are generally designed to discover patterns within a dataset. Whereas supervised methods might be used to predict the output variable from a given input, *unsupervised* methods might be tasked with discovering the relationships that allow for those predictions. Data clustering is one common example of an unsupervised method. As a toy example, it might be known that there are three types of fish in a population, but the only information available about the population is the size and color of the fish. One might be interested to know whether the fish can be grouped by similarities in their size and color patterns, and moreover, whether this correlates well with fish type. If size and color are the identifiable characteristics that determine fish type, then grouping fish by their similarities along these two variables should allow should allow for both fish type prediction and for a description of how color and shape vary across those groups. Example algorithms that can accomplish this task are k-means clustering, non-negative matrix factorization (NMF), and gaussian mixture models (GMM).

The second major type of unsupervised learning is dimensionality reduction, or methods which find reduced representations of the data that maintain its most important features. These include algorithms like Principal Component Analysis (PCA) or neural network derived autoencoders, which I will describe in greater detail below. Feature selection is one of the *most* important steps in building any statistical model, as we want to use the most important and informative features of our data as input variables and reduce noise from less informative features. High dimensional data sets associated with genomic data are particularly dependent upon effective methods for feature selection because each of the roughly 20,000 genes present in a given dataset represents a potential input variable, and many of these genes encode redundant information relative to the biological output variable of interest. Additionally, dimensionality reduction is useful for data visualization, because, as the renowned biologist Mike Wigler once exclaimed, we do not have robot eyes. Or, more technically, visually recognizing structural patterns within very high dimensional spaces is impossible, whereas lower dimensional embeddings of the data into 2D or 3D representations allow us to more easily assess emergent patterns that describe underlying relationships. A timely example of this in genomic biology is in the visualization of single cell derived data, which often consist of gene count matrices involving thousands of genes across thousands of cells that might each derive from dozens to hundreds of samples. In single cell analysis, visualization methods such as uniform manifold approximation & projection (UMAP) or t-distributed stochastic neighbor embedding (tSNE), have become standard methods to visualize relationships between cells in a population.

Together, supervised and unsupervised analysis encapsulates the majority of analyses employed in genomic biological research. The work described in this dissertation relied heavily on multiple methods from each category. Where appropriate to the data, I developed and employed several methods including: dimensionality reduction techniques, parameter estimation techniques, clustering methods, and multiple types of classifiers. Next, I will introduce each of the main methods I employed to complete the work in this dissertation. The dimensionality reduction methods included: principal components analysis (PCA) (Bishop 2006), a non-standard application of weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath 2008), and neural network (NN) autoencoders (Kramer 1991). Expectation Maximization (EM) was employed for parameter estimation (Bishop 2006). Of the unsupervised methods used in this work, the clustering algorithms included: non-negative matrix factorization (NMF) (Lee and Seung 1999) and gaussian mixture models (GMMs) (Bishop 2006). Finally, supervised analyses using feedforward multilayer neural networks (NNs) (Bishop 2006). I will describe how each of these algorithms work, in principle, and present simple examples of how they can be applied. For the data analysis methods developed as part of this thesis work, these algorithms were deployed through published libraries available in R or python and will be cited as they are used in the text.

1.4.2 - Expectation maximization (EM), a parameter estimation method

One hallmark of many algorithms used in this work is the existence of latent variables: variables that cannot be measured directly, but may be inferred through their effects on the variables that can be measured. One method commonly used for latent variable estimation is expectation maximization (EM). The EM algorithm features prominently in the small-RNA-seq analysis method I developed, TEsmall (O'Neill et al. 2018), which will be the focus of Chapter 2 of this thesis. A more complicated version of EM can also be used for general latent variable optimization, and will be discussed later in the context of GMMs. However, the application of EM to sequencing data analysis provides a good context for understanding how EM can be used for estimating unknown parameters, and will be used for this introduction.

The basic concept of EM algorithms, regardless of application, is as follows: a variable is estimated with some lower bound during the expectation step, and then this lower bound is used to maximize another function, often the likelihood of a particular fit of the data to a model (maximization step). The new optimized parameters for the function are then used to recompute a new lower bound for the variable of interest. This process is iterated until convergence, or a limit of iterations set by the user.

The EM algorithm provides an ideal solution to solving the problem of multimapper reads from repetitive regions of the genome. Briefly, short reads data from genomic repeat regions could map to multiple genomic loci, often with equally likely mapping scores. However, these genomic repeats are often interspersed with uniquely mappable regions. Thus, an mRNA transcript from a genomic repeat region will be likely to produce both uniquely mappable reads as well as reads that could align to other genomic loci (multimappers). This mixture of information from unique reads and multimapper reads can be leveraged to calculate the probability that a genomic repeat region is expressed and, as a corollary, the probability that a read was generated from that genomic locus. This type of problem is well-suited to EM algorithms, which were used for a transcriptome analysis pipeline in the lab called TEtranscripts (Jin et al. 2015). In the case of TEtranscripts, and the related software package TEsmall, the variable of interest for the expectation step is the probability that a short RNA read derives from a particular genomic location. The function that the TEtranscripts EM algorithm uses for the maximization step is the total sum of reads across a particular annotated region (eg., a gene or TE locus). The lower bound is initiated by the fractional distribution of a multimapper read to each of its potential mapping sites (1/N) (expectation step 0). This is a lower bound because a 1/N flat distribution will not contribute more than a single count to any potential locus of origin and allows for each read to be weighted by the relative certainty that it derives from a particular locus. This initial 1/N distribution of reads is used to estimate the relative abundances of all transcripts in a dataset and sets the initial estimates for the gene/TE read count table (maximization step 0). On all subsequent steps of the EM algorithm, the latest estimate of relative transcript abundances are used to update the weighted probabilities that a particular read derives from that particular transcript. The algorithm then continues updating both the read weights at a particular transcript locus (E step) and the transcript abundances as a sum of all associated read weights (M step) until the values converge. More specifically, once the estimated transcript abundances do not change by more than 10% in subsequent steps, the EM algorithm stops and outputs a final gene expression table for all gene and TE transcripts.

This application of EM algorithms for TEtranscripts/TEsmall is a relatively simple case in which the only latent variable is the true genomic origin of each multimapper read, and the relationship between read assignments and transcript abundance is clear -- that is, the relationship between the estimated variable in the E step (read weight) and the function in the M step (transcript abundance as a sum of read weights) does not need to be inferred. As such, the maximization is actually a minimization of the distance between the distribution of reads to each of the potential genomic locations it can map to and the other reads mapping to those locations. Now, let's expand this idea to the latent variable problem inherent in GMMs, where the latent variable is a mixture probability, or the probability that a sample belongs to a particular cluster (and that a particular clustering of the data is optimal). This latent variable model, as addressed by EM, can be described as $p(x;\theta) = \sum_z p(x,z;\theta)$ and the log likelihood of the latent variable model as $l(\theta) = \sum_{i=1}^{n} log \sum_{z^{(i)}} p(x^i, z^i; \theta)$. This is a complicated likelihood to maximize, given that we do not initially know how z is distributed, and there are many unknown parameters. Therefore, we will employ a technique common in machine learning and Muay Thai (Workshop 2019), and use an ELBO, or Evidence Lower Bound, to create a clever function known to be less than the true target distribution using Jensen's inequality. This evidence based lower bound (ELBO) on the log-likelihood of the data can then be maximized using EM. This derivation is taken from Tengyu Ma and Andrew Ng's CS299 course, and is available at (http://cs229.stanford.edu/notes2021fall/cs229-notes8.pdf). I will not include the full derivation, but rather describe how this trick allows us to successfully perform EM on hidden variables without known distributions. We allow Q to be a distribution over z where z is non-zero, and allow the log likelihood to remain $\log p(x;\theta) = log \sum_z p(x,z;\theta) = log \sum_z Q_z \frac{p(x,z;\theta)}{Q_z} \geq \sum_z Q_z \, log \frac{p(x,z;\theta)}{Q_z}$, where f(x) = log(x) is a concave function and Jensen's inequality can be applied. We can now set the probability distribution Q to be any function that is convenient for solving our EM problem, since any function Q will satisfy the inequality in the above equation. We choose the posterior distribution of z given our data and the parameters of the distributions from which they are drawn $Q(z) = p(z|x;\theta)$. When Q is substituted in the equations above, it returns to the original log-likelihood function. This property of an ELBO is often written as ELBO(x^i;Q_i;θ) and is necessarily less than the log-likelihood of $p(x)$. In this framework, the expectation step involves calculating $Q(z) := p(z|x;\theta)$ over all samples, and then the maximization step calculates the argmax of θ ELBO(x^i;Q_i;θ), summed over all samples. This generalization of EM will converge monotonically and result in a maximized likelihood of $\log p(x;\theta)$.

1.4.3 – Principal components analysis (PCA)

While applying each of the algorithms to biological data in the course of this thesis, it became clear that data pre-processing was just as important to the eventual success of the projects as algorithm selection. I will thus briefly describe the relevant dimensionality reduction methods I employed including: PCA, WGCNA, and autoencoders. Additionally, while PCA is an initial step for many other dimensionality reduction algorithms, it can also be used on its own for data visualization and other applications.

PCA, or principal components analysis, is a linear transformation of a dataset into a series of unit vectors that are uncorrelated with respect to each other. These are calculated by rotating or changing the basis of a dataset to align with the axes of largest variance within the data, while minimizing information loss. This transformation is how PCA achieves its aim of dimensionality reduction, as the relevant variance is encoded along a finite set of new basis vectors, or principal components, with each subsequent component describing a smaller source of variance, not previously described by any lower component of the data. While PCA-based vectors are conventionally defined to be orthogonal to one another allowing for convenient shortcuts, such as allowing the user to choose to work with the set of principal components that explain a given amount of the variance in a dataset (say 90%) rather than working with a fixed number of principal components that may account for different fractions of the variance in different data contexts. This transformation can be written explicitly as T=XW, where W represents the rotation, X is the original data matrix, and T is the reduced data matrix of principal components. When a limited number of principal components are kept, for the purpose of dimensionality reduction, this is represented as $T_L = XW_L$. The squared error between X and X_L can be minimized to determine the number of components which must be kept to faithfully preserve the data.

1.4.4 – Weighted gene correlation network analysis (WGCNA)

WGCNA was not originally developed as a form of dimensionality reduction; rather it was developed as a way to characterize networks of co-expressed genes in microarray-based gene expression data (Langfelder and Horvath 2008). WGCNA functions on the biological principle that genes function within non-random networks, and that these networks are regulated by orchestrated cohorts of transcription factors. In these gene expression networks, genes that are regulated by the same sets of transcription networks will show correlated expression values across a wide variety of samples and conditions. This can be represented statistically as a scale free network, in which the density of correlated genes follows a power law, represented by $P(k) \sim k^{-\gamma}$. In this network description, each gene is a node that is connected to another gene node by an edge if those two genes are co-expressed across many samples. P(k) is the fraction of genes that are co-expressed with k other genes. The parameter, γ, describes the relative size of correlated gene modules. The larger the γ parameter, the fewer co-expressed genes are found in each hub of the network. In WGCNA, a matrix of correlations between all genes are calculated pairwise, and represents the similarity of each gene to one another. This can be calculated in a signed or an unsigned manner where the former denotes if two genes are positively or negatively correlated with each other while the latter simply states their correlation value, regardless of direction, $s_{ij}^{signed} = 0.5 + 0.5cor(x_{i,}x_j)$. This similarity matrix is then confined into a network structure by an adjacency matrix $a_{ij} = s_{ij}{}^{\beta}$, which describes how sets of genes are connected to each other by a threshold β. This threshold term is chosen by the user as the minimal integer value which satisfies the scale free network assumption. After genes are grouped into hubs by this adjacency matrix, known in WGCNA as modules, these modules are represented by a single value called an eigengene. An eigengene is the value of the first principal component of all of the genes within a module, and makes WGCNA a robust way to represent trends of co-expression within a set of genes. WGCNA is a biologically informed way to reduce the number of features being used as inputs in a downstream statistical model, and typically reduces an expression matrix of ~20,000 genes to ~40 eigengenes.

1.4.5 – Autoencoders, a neural network based method for dimensionality reduction

Before discussing autoencoders themselves, I must first introduce neural networks (NNs), the underlying architecture of autoencoders. A neural network is an extremely flexible type of model that has been deployed to solve a large variety of statistical tasks. NNs can be grossly defined by their ability to learn an underlying representation of a given dataset that can be used to transform a given input data point into a limited choice of outputs. For example, given sets of images of animals (cats and dogs) as input, a NN might be trained to label which type of animal is present in a given image (cat or dog). NNs are built up from layers of computational "neurons", each representing a particular mathematical function, called an activation function, that will be applied to any incoming signal (input data). Therefore, a NN is really just a mathematical description of how multiple units, characterized by their respective activation functions, can be linked together to create complex data transformations. The most commonly used activation function in modern NNs is the Rectified Linear Unit (ReLU). ReLU is a non-linear transformation described programmatically as max(0,z) and is widely employed because of the facility of computing its derivative. Derivative calculations for these activation functions are critically important in the context of training neural networks, as will be discussed later.

In a NN, nodes are organized into layers beginning with the input data nodes, and where outputs from each layer are designed to connect as inputs to subsequent layers according to the rules of a given NN architecture. Users may choose, in designing a particular neural net architecture, how many nodes are present in each layer and how each layer of nodes is connected to its input and to the next output layer. The weights of these connections are subsequently tuned by the algorithm itself during the NN training phase to achieve optimal performance on a given task – which typically involves learning how to correctly assign a label to a particular piece of input data (e.g., assigning the label “cat” to an input image of a cat). This is typically expressed as a loss function between the known data labels and the calculated NN output labels at each epoch of the NN training phase.

The wide variation with which NN nodes can be connected together leads to incredible flexibility and diversity as to what tasks a neural network can be trained to perform. The cost of arbitrarily complicated NN architectures comes during attempts to optimize the model, as not all NN architectures can be easily optimized for a given task. Specifically, a NN is initialized with a randomized set of weights on each NN node. It then enters an iterative training phase that alternates between: (1) a feedforward computation of the NN output from the current set of node weights, and (2) an update step which is achieved through a backward propagation of the errors to the weights in each prior layer. In each of these iterative steps between feedforward computations and error backpropagation, the algorithm attempts to tune each node weight in order to better match the input data to the output labels. The contribution of each node weight to the error between the predicted output (as calculated by the NN) and the actual label of the output data is found by calculating the derivative of the error function with respect to each of the NN’s weights. This allows for changing each node weight in a way that is proportional to its contribution to overall error. This is a particular application of a general strategy for optimization called gradient descent. The original input data is then fed again to the updated model with newly calculated weights until the prediction errors (difference in the value from the loss function) from one epoch to the next has stabilized. Immense research has been put into developing gradient descent algorithms, since these form the heart of NN optimization strategies as well as many other machine learning strategies.

Autoencoders are a form of neural network with a bottleneck structure that allows them to serve as a dimensionality reduction method. In contrast to the NN architectures described above, autoencoders do not try to classify data into a set of labeled outputs. Instead, autoencoders are designed to find a reduced representation of a dataset that can fully recapitulate the most salient information of that original data. In other words, autoencoders learn how to both encode and then decode the data, with the eventual output looking as close as possible to the original input. In the simplest case, the loss function of an autoencoder to be optimized is the mean squared error between the original input vector, and the output that has been passed through an inner layer with fewer dimensions. The minimization of the loss between the input and compressed output forces the activation weights of the model to encode the structure of the data representing a lower dimensional manifold. One can access this manifold by plotting the activation weights, or the outputs from the transformation associated with each node, from the inner bottleneck layer. You will see an example application of autoencoders in Chapter 4.

1.4.6 – Non-negative matrix factorization (NMF), an unsupervised learning method

As introduced above, unsupervised learning methods can be used to discover patterns within a dataset, when examples of labeled patterns are not available. One major type of unsupervised learning method employed in this work involved non-negative matrix factorization (NMF) (Lee and Seung 1999). NMF is a form of matrix decomposition in which a matrix V_{ij} is factorized into two submatricies W_{ik} and H_{kj} which, when multiplied together, will approximate the original matrix V_{ij}, where *i* and *j* represent genes and samples respectively. This is generally achieved through random initialization of the matrices W and H, and a minimization of the least squares error between V and the product of W and H by a form of gradient descent, although other update strategies exist. In the implementation of SAKE, a version of gradient descent called coordinate descent was used to update the W and H matrices in the optimization protocol and is described in Ho et al. (Ho et al. 2018). This works as a clustering method, because the probability of assignment of a sample to a cluster *k* is obtained by $m(\Theta, j) = \frac{H(\Theta, j)}{\sum_{j=1}^{k} H(\Theta, j)}$, from the H matrix. Similarly, using this method one can assess the importance of a gene to a particular cluster *k* by performing the same computation across matrix W for all *i* to return the probability a gene contributes to a particular cluster *k.* These computed probabilities of cluster membership and gene importance highlight the merits of using NMF as a clustering algorithm. Often in genomic analysis, not only are we interested in assigning our samples into groups, but we are also interested in what genes contributed to a sample's participation in each of those groups. This is what allowed my lab to originally associate known ALS pathways with each of the molecular subtypes discovered in our ALS patient samples. However, this method is limited in that it is a *de novo* clustering algorithm, and cannot classify new samples without rerunning the entire pattern finding method on all samples. This can be problematic, as we are often interested in looking for the signature of these subtypes in new data, and is one of the main challenges addressed by this thesis.

1.4.7 – Gaussian mixture models (GMMs), an unsupervised learning method

Gaussian mixture models (GMMs) are another form of unsupervised clustering algorithm that can be used to detect latent structure in data. GMMs expect that there are a finite number of groups in a dataset, each of the which are normally distributed in each of the dimensions of your sample features, potentially genes. GMMs go beyond simply fitting the data to gaussian distributions; these also attempt to find latent variables within the data that allow for clustering, the mixture probabilities. These mixture probabilities estimate the likelihood that a sample belongs to a particular gaussian distribution (one of the clusters), which is the main result we are interested in. Latent variable problems like these are generally addressed by an expectation-maximization (EM) algorithm, a particular application of maximum likelihood frameworks that was introduced in Chapter 1.4.2.

Mathematically, GMMs are described by a joint probability distribution of a multinomial of multiple gaussian distributions. There are three parameters of the model ϕ,μ, and Σ, which are k dimensional vectors representing the mixture weights of each gaussian, the means of each gaussian, and the covariance matrices of each gaussian, respectively. The likelihood function is given by $l(\phi, \mu, \Sigma) = \sum_{i=1}^{n} log \sum_{z^i}^{k} p(x^i | z^i; \mu, \Sigma) p(z^i; \phi)$, where *z* represents the probability of participation of a sample to a particular gaussian *j*. It is impossible to differentiate this likelihood function, so we will use EM as described above to overcome the hidden latent variable *z* and guess by flipping between calculating the posterior distribution with respect to the latent variable and the likelihood. We will calculate the expectation step of the posterior as $w_j^i := p\left(z^i = j \middle| x^i; \phi, \mu, \Sigma\right)$ then update each of the unknown parameters as $\phi_j := \frac{1}{n} \sum_{i=1}^{n} w_j^i$, $\mu_j := \frac{\sum_{i=1}^{n} w_j^i x^i}{\sum_{i=1}^{n} w_j^i}$, $\Sigma_j := \frac{\sum_{i=1}^{n} w_j^i (x^i - \mu_j)(x^i - \mu_j)^T}{\sum_{i=1}^{n} w_j^i}$. The output of this algorithm would be a description of the k clusters inferred to be present in the data, the likelihood that each sample (n) belongs to one of those k clusters, and the means and variances of the underlying data (e.g., gene expression representations) that differentiate each of the k clusters.

1.4.8 – Feed-forward multilayer perceptrons, a neural network (NN) for classification

Artificial neural networks (ANNs), their general architecture, and common biological applications were discussed in detail in section 1.4.5. Rather than reintroduce these concepts, I will describe the basic neural network architecture that was used in Chapter 4 of this thesis: a feed-forward multilayer perceptron (Bishop 2006). As described above, the basic architecture includes: a set of input data nodes, multiple hidden layers of fully connected nodes, as well as an output layer. In general, the inner layers are hidden with their respective activation functions. The output layer is the layer constrained to have a sigmoidal, or softmax, activation function. This output layer gives the probability that a given input data sample belongs to one of a finite number of output classes. These models are trained by minimizing the loss function of cross entropy, or log loss, between the probability of the true class of a sample ($p = 1$) and the predicted probability for that class for a sample (e.g. $p=0.7$). Of note, these multifactor classification problems use one-hot encodings of a particular class label as outputs so that this cross-entropy paradigm can be used where the true class is assigned a probability of 1 and all other classes a probability of 0. This allows loss to be represented evenly with respect to each class.

Like many “black box” learning methods, it can be difficult to decipher what exactly a NN is learning, when trained to transform a given set of input data into known output layers. One would hope that the hidden inner layers represent a non-linear transformation of the data into a simpler and more informative data space. Saliency analysis can be used to discover some characteristics of how trained NNs are representing the data. This may be accomplished by removing some input data and viewing how this changes the outputs, for example. Alternately, ablating some inner nodes to decipher how these contribute to the overall output can also be used. Saliency analysis can also be used to ensure that NNs are not learning technical artefacts unrelated to the biological problem (Koo and Eddy 2019).

Neural networks are becoming increasingly popular in the analysis of biological data as we create large -omic datasets like those generated by chromatin immunoprecipitation (ChIP-seq), Assay for Transposase-Accessible Chromatin (ATAC-seq) or Hi-C a chromosomal conformation capture method. Neural networks built from this data are being applied to predict variables like: DNA and RNA binding site specificity of proteins, gene expression levels, and methylation status of DNA and are described in a detailed review by Zou et al. (Zou et al. 2019). However, in this thesis neural networks will be used to look for transcriptional structure in bulk RNA-seq data, and assess the landscape of gene regulatory networks as they present in the heterogeneous disease ALS. These neural networks will be applied in both an *unsupervised* manner as described in section 1.4.5 in the form of an autoencoder, and in a *supervised* manner in which previous molecular subtype cluster assignments determined by the de-novo clustering algorithm NMF will be used as training labels. A neural network classifier will be trained to accurately predict these labels on new incoming data in an example of converting an unsupervised problem into a supervised problem. This classification algorithm will then be used to predict subtype on a novel dataset.

Chapter 2: Description of current available methods for computational genomic analysis of transposon biology

2.1 An introduction to a review of TE specific computational tools and techniques

As described in the previous chapter, analysis of transposon biology requires special care as transposons are highly repetitive in nature and difficult to disambiguate in computational analyses. This uncertainty results in TE associated reads widely being discarded in traditional analysis of sequencing based datasets. This chapter is a reformatted version of a review addressing this topic published February 2020 in Philosophical Transactions of the Royal Society B, a special issue on mobile genomics. Here I address a number of the computational challenges faced by the field of transposon biology in analysis of sequencing based "-omic" data, not including genome assembly (genomics, transcriptomics, and other genome-wide functional genomics data). I was the primary author of the publication, and was responsible for writing the article and the production of all figures with the exception of Figure 4, and the section about single cell technologies for transposons. In addition, this chapter includes a brief segue into the next chapter about TE-aware small RNA biology and software.

2.2.1 The Mobile Genomics: Tools and Techniques for Tackling Transposons Manuscript

Mobile Genomics: Tools and Techniques for Tackling Transposons

Kathryn O'Neill[1], David Brocks[2], and Molly Gale Hammell[1]

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA.

[2] The Weizmann Institute of Science, Department of Computer Science and Applied Mathematics, Rehovot, IL

Abstract

Next generation sequencing (NGS) approaches have fundamentally changed the types of questions that can be asked about gene function and regulation. With the goal of approaching truly genome-wide quantifications of all the interaction partners and downstream effects of particular genes, these quantitative assays have allowed for an unprecedented level of detail in exploring biological interactions. However, many challenges remain in our ability to accurately describe and quantify the interactions that take place in those hard to reach and extremely repetitive regions of our genome comprised mostly of transposable elements (TEs). Tools dedicated to TE derived sequences have lagged behind, making the inclusion of these sequences in genome-wide analyses difficult. Recent improvements, both computational and experimental, allow for the better inclusion of TE sequences in genomic assays and a renewed appreciation for the importance of TE biology. This review will discuss the recent improvements that have been made in the computational analysis of TE derived sequences as well as the areas where such analysis still proves difficult.

Keywords

Transposable elements; Computational Genomics; Retrotransposons; Single cell Analysis

Introduction

While several types of genomic repeated sequences exist, the largest fraction of the human genome, approximately half, is comprised of transposable elements (TEs)(Pace II and Feschotte 2007), though some groups estimate much larger TE fractions(de Koning et al. 2011). These TEs, often called transposons or jumping genes, are DNA sequences that have, or once had, the ability to mobilize within the genome, either directly or through an RNA intermediate. TEs are present, to varying degrees, in the genomes of all known types of organisms, both prokaryotic and eukaryotic, with some species showing more genomic transposons than host sequences(International Rice Genome Sequencing Project et al. 2005). Several excellent reviews have discussed the many and varied types of TEs(Feschotte and Pritham 2007; Slotkin and Martienssen 2007; Levin and Moran 2011). Briefly, TEs come in two major types. Class I TEs, also called retrotransposons, first transcribe an RNA copy that is then reverse transcribed to cDNA before inserting elsewhere in the genome. Class II TEs, also called DNA transposons, directly excise themselves from one location before reinsertion. In the human genome, the vast majority of TEs are of the Class I, retrotransposon type. Nearly all human TEs have lost the ability to fully mobilize (Boissinot et al. 2000; Sheen et al. 2000; Brouha et al. 2003), with the human specific LINE-1 element (L1HS) being the only fully autonomous TE with the ability to generate new transposition events to date. However, most TEs have retained some level of functionality, including the ability to direct their own transcription. Thus, transcriptome-wide sequencing assays, like RNA-seq, frequently include transposon-derived transcripts among the set of expressed sequences. Moreover, some transposon transcripts have been coopted to play a role in host function, particularly during early development, such that some expressed transposon transcripts have been shown to be necessary for proper cell differentiation and maintenance of identity (Lu et al. 2014; Grow et al. 2015; Jachowicz et al. 2017; Rowe et al. 2013; Percharde et al. 2018). In addition to their roles in general cellular function, several types of transposons have become intricately entangled within gene regulatory networks (Chuong et al. 2017), contributing both to cis regulatory sequences (Imbeault et al. 2017; Chuong et al. 2016; Sundaram et al. 2014) as well as general chromatin environments (Pontis et al. 2019; Venuto and Bourque 2018; Raviram et al. 2018). For this reason, it is paramount we consider the contribution of repetitive elements as we unravel the genomic and epigenomic landscapes that control gene expression.

Properly accounting for repetitive regions in most genomics analysis settings requires special considerations for the challenges presented by the number of nearly-identical transposon sequences dispersed throughout our genomes. Thus, reads derived from these regions are frequently discarded in most sequencing data analysis protocols due to the difficulty in properly assigning TE-derived reads to the correct locus of origin. Few packages explicitly support inclusion of repeats and some intentionally discard reads from these regions, as discussed in a recent review(Goerner-potvin and Bourque). Of the packages designed to address TEs, many tools focus on the detection of novel TE insertions or TE-associated genomic rearrangements. Few tools are developed specifically to address regulatory and transcriptional activity of TEs in common assays, such as RNA-seq, chromatin immunoprecipitation sequencing (ChIP-seq), cross-linking immunoprecipitation sequencing (CLIP-seq), and small RNA-seq (sRNA-seq). In this review, we seek to provide an overview of the packages that explicitly support the inclusion of TE sequences in differential expression and binding analyses, and the strides which have been made to improve our ability to resolve ambiguously mapped reads in genomics analysis.

Annotation and de novo detection

A well assembled and annotated genome is the foundation for effective analysis, as all subsequent analyses discussed below require a reference genome as well as a map of gene and TE positions. While many genomes have near complete assemblies, and extensive annotation, the quality of both tends to drop over repeat rich regions for the same reasons discussed above: ambiguity in placing near-identical sequence reads from highly similar copies of related transposons. This ambiguity leads to non-contiguous and erroneous chromosomal assembly, which will feed forward into any genomics analyses using these assemblies(Treangen and Salzberg 2012). Genome assembly has benefitted immensely from long read sequencing technologies, particularly in the context of highly repetitive centromeric regions and in nested repeating elements(Jain et al. 2018; Gordon et al. 2016). While these long read technologies are improving the reference genomes used to map new datasets, one caveat is that transposons are often polymorphic within populations, such that each new sample sequenced is expected to have many non-reference transposon-associated insertions, deletions, and other structural variants that may be rare or private(Wong et al. 2018; Chaisson et al. 2019).

Once a high-quality assembly is constructed, the process of annotation may begin. Many curated annotation databases have been developed for identifying repeat elements. For an in depth review of annotation practices and existing repositories please refer to the review by Goerner-Potvin et al.(Goerner-potvin and Bourque) Here the distinction between TE-, genome-, and polymorphism-focused annotation repositories is emphasized in addition to a list of software for de novo insertion detection. The most widely used database of TE consensus sequences is RepBase (Bao et al. 2015), which provides the sequences with which genome-specific annotation files are constructed. These annotation files are available through the University of California Santa Cruz Genome Browser (UCSC) and

RepeatMasker (Kent et al. 2002). While new RepBase consensus sequences require a subscription, several open databases for repeat annotation are available in addition to UCSC including: RepetDB (Amselem et al. 2019), ERVdb (Gifford et al. 2018), Dfam (Hubley et al. 2016), TREP (Wicker et al. 2002), SPTEdb (Yi et al. 2018a), ConTEdb (Yi et al. 2018b), and mips-REdat (Nussbaumer et al. 2013). The ideal database for analysis will vary depending on model organism and transposable elements of interest, as some databases are species and TE type specific.

Mapping

After the construction of a well annotated reference genome, one is faced with the task of mapping experimental data to the appropriate reference. Even with a perfectly annotated and constructed genome, ambiguously mapped sequencing reads still present a challenging problem. One of the first approaches to address this problem, designed for RNA-seq analysis, was to probabilistically assign multi-mapped reads to regions that also show a higher density of uniquely mapped reads, i.e. reads with a single best genomic alignment under the mapping software's heuristics (Mortazavi et al. 2008). However, this was a highly gene centric model that was primarily focused on host gene expression, and was not explicitly intended for estimating expression from TE loci. Moreover, this approach is biased toward regions that have some uniquely mappable content. Unfortunately, the most recently integrated TE insertions are also the least likely to be uniquely mappable, and are thus the most likely to be lost or underestimated by these methods. To highlight this, Figure 1 displays the estimated mappability of several different types of TEs in the human genome, with a specific emphasis on younger types of TEs shown to be active in the human genome (Mills et al. 2007). Mappability in this plot was defined as the inverse of the number of times a simulated 76bp paired end read mapped to the genome, allowing three mismatches. Mappability was scored per nucleotide with the score assigned to the first nucleotide of the read. This track was procured from an in-depth analysis performed by Sexton and Han which considers the many parameters that contribute to the mappability of a particular sequence including the mapping software chosen and the length of the sequenced read (Sexton and Han 2019). These analyses still return to the same basic theme displayed in Figure 1: mappability rates vary for different types of transposons, and the most recently inserted transposons are the most likely to be discarded by standard analyses that rely on uniquely mapped reads. In other words, the transposons that present the most problems in genomics analyses are precisely those that are more likely to be functional in terms of: carrying fully functional promoters, encoding for functional proteins, and, rarely, mobilizing within the genome. In addition, many older elements with degraded versions of these components have been recycled to play roles in cis regulatory architecture (Feschotte 2008).

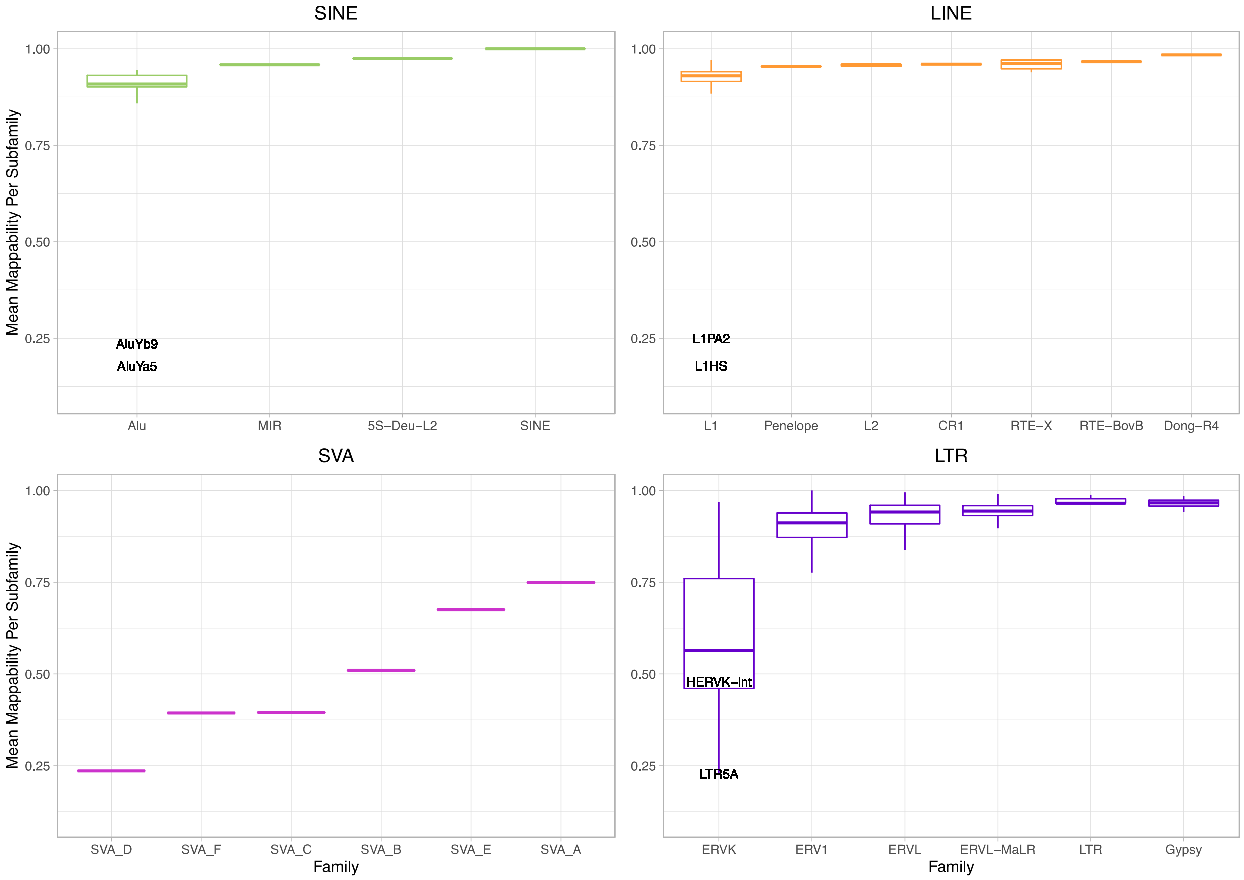


Figure 1 Estimated mean mappability for different types of TEs in the human genome

Mappability tracks from the analysis by Sexton and Han for hg38 were used to construct mean mappability estimates (average probability that a pair of 76bp reads would map uniquely to a genomic instance of that TE). These were then aggregated by subfamily (L1HS is a human specific subfamily of the LINE class). Some TEs have accumulated enough mutations across each locus that nearly all copies are uniquely mappable. Very recently inserted, and/or still active TEs, show the lowest mappability rates with many copies still very close to the consensus sequence (e.g. Alu and SVA types). In contrast, many older SINE and LINE TEs have high mappability rates and can easily be assessed using only uniquely aligning reads with standard analysis procedures. Mappability was calculated by counting number of times a 76bp paired end read (242-mer with an internal gap of 100 nt) would map within the genome at a particular nucleotide where that nucleotide was the beginning of a 242-mer.

Most genome alignment software is aware of the difficulties posed by ambiguously mapped reads, and thus provide extensive parameter sets designed to allow the user to choose the number of alignments considered for each sequenced read. This includes standard genome mapping software applicable to genome resequencing studies as well as ChIP-seq based studies of protein-DNA binding, such as BWA(Li and Durbin 2010), bowtie (Langmead 2010), and Novoalign (http://novocraft.com/). For RNA-seq aligners there are two approaches, those that align to reference transcriptomes and those that align to genomes. Transcriptome methods like kallisto (Bray et al. 2016) and salmon (Patro et al. 2017) perform pseudoalignments with transcript derived k-mers and can attempt to build the reference transcriptome from the RNA-seq data itself. Salmon can be specified to report unmapped reads, kallisto does not include this option. While pseudoalignment is very fast, computationally less intensive, and helpful in organisms without a reference genome, it can be complicated in the context of repetitive elements, where all of the caveats that make genome assembly difficult (discussed above) would also apply to de novo transcriptome assembly. With regard to genome based RNA-seq aligners, there are a number of packages available including: STAR (Dobin et al. 2013), HISAT2 (Kim et al. 2019), GSNAP (Wu et al. 2016), Novoalign, RUM (Grant et al. 2011), Minimap2 (Li 2018) and others (Baruzzo et al. 2017). In the context of sRNA-seq data, short read genome-based aligners (BWA (Li and Durbin 2010), bowtie (Langmead 2010) and SCRAM (Fletcher et al. 2018)) that do not consider splice junctions tend to work as well or better than RNA-seq tailored algorithms, with SCRAM being specifically designed for small RNA analysis pipelines. Another approach to improve mappability would be to incorporate long-read sequencing methods, as longer reads contain more information and can serve as a way to reduce ambiguity in the context of RNA-seq. Many of the previous aligners like STAR, HISAT2 and GSNAP have been applied to long-read sequencing data after error correction (Krizanovic et al. 2018) and have been shown to work well. In addition, algorithms like BLASR (Chaisson and Tesler 2012), GraphMap (Sovic et al. 2016), rHAT (Liu et al. 2015), LAMSA (Liu et al. 2017), Kart (Lin and Hsu 2017), NGLMR (Sedlazeck et al. 2018), and lordFAST (Haghshenas et al. 2019) have been developed specifically to address the increased length and error rates associated with long read technologies.

Some tools designed to improve mapping rates for repetitive regions work after an initial analysis with one of the tools listed above. These standalone tools can use alignment files as input and then attempt to statistically redistribute the ambiguous reads based on distributions of neighboring alignments. One such algorithm is MMR (Kahles et al. 2016) which iteratively redistributes ambiguously mapped reads across their respective loci to maximize smoothness of multimapped read distribution in the context of unique reads, or reduce the variance in coverage. Another is a Gibbs sampling method (Wang et al. 2010) which uses stochastic redistribution of multimapped reads, normalized to the background distribution, in order to iteratively search for the most likely locus of origin. This type of iterative statistical technique for optimal assignment of reads to the correct loci has been picked up and elaborated on by several different groups, and represents a theme throughout the review. While it does not employ the statistical redistribution of reads, CoCo (Deschamps-Francoeur et al. 2019) is a package which corrects and salvages multimapped reads by taking into consideration nested genomic architecture, a common feature associated with TEs.

Analysis

The next step in a general NGS sequencing analysis pipeline is to annotate and quantify those reads which mapped to the genome. The mapping profiles will vary widely based on molecular context of the sequencing library. Each type of NGS data comes with its own challenges in the context of highly repetitive elements. The remaining sections will go through analysis strategies for each of the most common NGS data types in detail. The tools in these sections are listed for reference on Figure 2, where they are grouped by the experimental assays used to generate the data. Table 1 gives references and links to the software for all tools described.

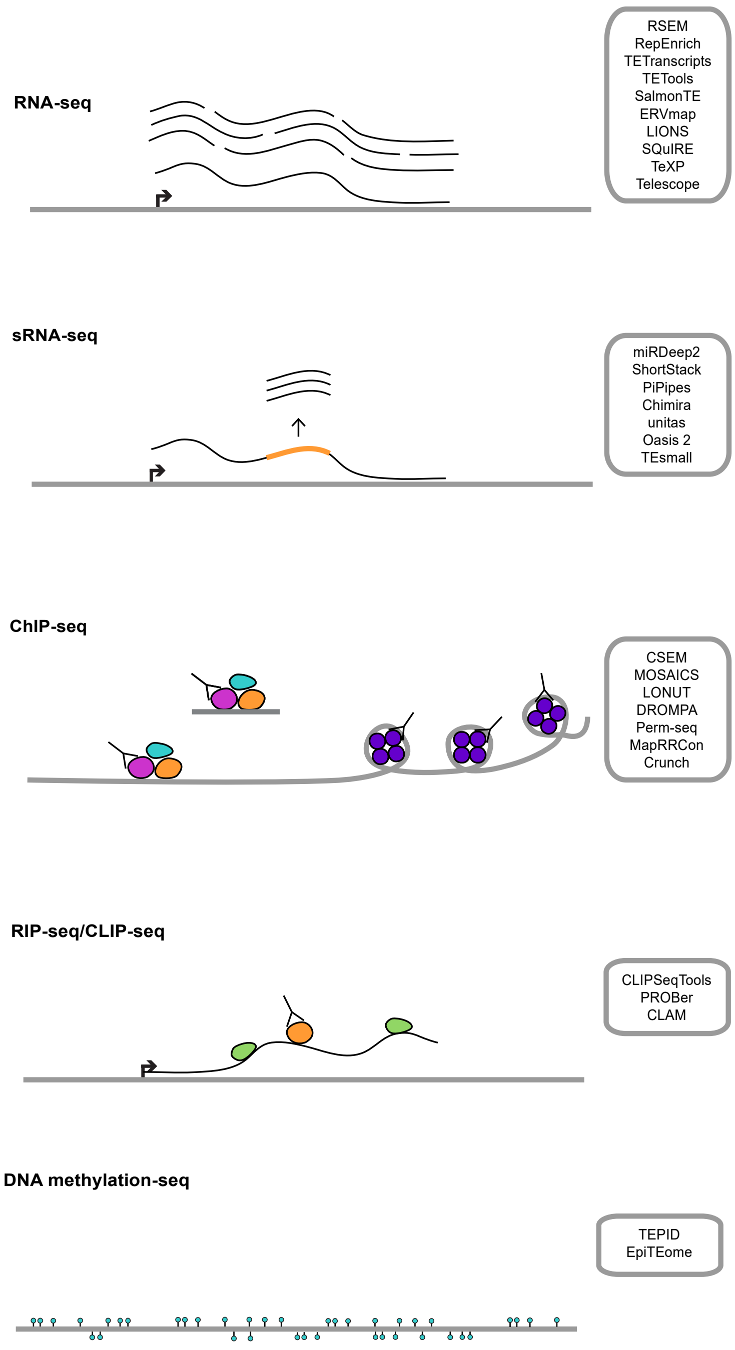


Figure 2 Published tools available for including repetitive regions in several common genomics analysis protocols.

These have been divided into those that are geared toward RNA expression analysis (RNA-seq), small RNA expression analysis (sRNA-seq), genome and chromatin binding factors (ChIP-seq), RNA-binding factors (RIP/CLIP-seq), and DNA methylation analysis (DNA methylation-seq). A table describing these tools (Table 1) also provides links and references for the software and associated publications.

RNA-seq

RNA-seq for expression analysis is one of the most well studied areas in genomics, and this is also reflected in the diversity of tools available for analysis of transcripts from repetitive regions. RNA-seq data derived from short-read sequencing platforms is comprised of small fragments, derived from short single- or paired-end reads tiled across the region of a transcript of origin. Of the tools which have been developed to facilitate transcriptional analysis of repetitive elements, here we will focus on those which take into consideration ambiguously mapped reads. How to address ambiguously mapped reads is an old problem in genome science particularly when using older sequencing technologies from which reads were much shorter (~36 nt) than what we currently consider a short read (~150 nt). These early RNA-seq packages were largely gene centric, as investigation of repetitive elements with these earlier technologies was (and remains) a challenge. However, the basic principles for probabilistic redistribution of ambiguously mapped reads emerged at this time. The first strategies employed a single-step multimapped read redistribution based on the number of uniquely mapped reads at each locus (Mortazavi et al. 2008). This was followed quickly by an expectation maximization (EM) algorithm to iteratively estimate the most likely expression levels of gene transcripts based on relative counts of unique and multimapped reads (Li et al. 2010). In addition to probabilistic redistribution of reads, packages like Cufflinks (Trapnell et al. 2010) and HTseq (Anders et al. 2015) have multimapper modes where ambiguously mapped reads are weighted by the relative number of genomic alignments (as 1/n, where n is the number of potential alignments in the genome). The package Scavenger (Yang et al. 2019a) considers multimapped reads and uses an intermediate consensus assignment with remapping to rescue unmapped reads.

Differences in strategies used to address multimapped reads and their associated limitations are outlined in detail by Treangen and Salzberg (Treangen and Salzberg 2012).

As interest broadened to begin investigating transposon expression through RNA-seq explicitly, several packages were developed to handle transposons separately from the rest of the transcriptome. Among the first TE-centric packages was RepEnrich (Criscione et al. 2014) which functions by creating repetitive element pseudochromosomes, which are a series of contigs that represent all of the genomic instances of each transposon subfamily annotated in RepeatMasker, concatenated onto a single region. These subfamily pseudochromosomes were then used to identify reads that mapped only to one subfamily of transposons, such as the human specific LINE element L1Hs, even if the exact generating locus was still ambiguous. This was able to separate the level of uncertainty to finer detail, such that reads could be described as: unique in the genome, unique to a particular subfamily, or ambiguously mapping to multiple types of transposons. Similar to RepEnrich, TETools (Lerat et al. 2017) is another transcript quantification method which uses a detailed annotation file or 'rosette' to facilitate quantification from TE derived reads, and which again aggregates reads at the subfamily level. TeXP (Navarro et al. 2019) is a package which focuses on LINE-1 elements specifically and models spurious genome transcription to more accurately quantify LINE-1 expression. TEtranscripts (Jin et al. 2015) was the first TE-centric algorithm to implement statistical read redistribution to handle multimapped reads. TEtranscripts uses an expectation maximization algorithm to find the most likely distribution of ambiguously mapped TE-derived RNA-seq reads, and also includes expression estimates for both host genes and transposable elements in the output. After TEtranscripts, other packages have been developed to expand the methods used for statistical read redistribution including MMR (Kahles et al. 2016) and SalmonTE (Jeong et al. 2018), with SalmonTE being unique in its use of a pseudoalignment strategy from the authors of the original Salmon (Patro et al. 2017) package in order to bypass the mapping step typically used in RNA-seq analysis. Yanagi (Gunady et al. 2018) expands on this pseudoalignment strategy by mapping to a segmented version of the transcriptome to reduce ambiguity of mapping.

In the packages described above, quantification was performed at the subfamily level, as determining the specific expressed genomic loci within a subfamily is quite difficult for transposable elements that are close to the consensus sequence. However, several newer packages have been released to address the need for locus specific quantification of TE derived transcripts. TE-centric packages include SINEsFIND (Carnevali et al. 2017), and ERVmap (Tokuyama et al. 2018) which are specialized for their respective TE family of interest. Two pipelines used genome guided de novo transcriptome assembly with Trinity (Haas et al. 2013) to quantify TE expression at a locus specific level: TEcandidates (Valdebenito-Maturana and Riadi 2018) and a pipeline described by Guffanti et al. (Guffanti et al. 2018). More recently, SQuIRE (Yang et al. 2019b) (Software for Quantifying Interspersed Repeat Expansion), and Telescope (Bendall et al. 2019) adapted the EM-based read redistribution strategies described above to infer originating loci of ambiguously mapped reads, using uniquely mapped reads surrounding the locus to guide the EM read redistribution.

One of the motivating reasons to study transposable elements is for their influence over regulatory networks in our genome. To address this specifically, a final type of RNA-seq analysis package has been released at the interface of gene-centric and TE-centric models. LIONS (Babaian et al. 2019) is a novel package which detects novel fusion events that connect TE promoter sequences to downstream coding gene sequences. These chimeric

TE/gene transcripts represent one of the many ways that TE promoter elements might affect regulation of adjacent genes.

Small RNA-seq

Cells regulate transposable element expression using multiple strategies. The most potent silencers of TEs in germline cells are small RNAs (sRNAs) of the PIWI-interacting RNA (piRNA) class (Malone and Hannon 2009). In somatic tissues, two additional classes of small RNAs contribute to TE silencing: short interfering RNAs (siRNAs) derived from expressed transposon transcripts (Malone and Hannon 2009) and the more recently described 3′ tRNA derived fragments (3′ tRFs) (Schorn et al. 2017). Therefore, it is integral to the study of transposon biology to consider sRNAs and accurately quantify their production. To this end, several packages have been released to investigate sRNA species, which prove particularly challenging when derived from repetitive loci in the genome as they are short in length, typically between 18-36 nucleotides. Packages like MiRdeep2 (Friedländer et al. 2012), ShortStack (Axtell 2013), PiPipes (Han et al. 2015), Chimira (Vitsios and Enright 2015), sRNAtoolbox (Rueda et al. 2015), Oasis 2 (Rahman et al. 2018), and Manatee (Handzlik et al. 2019) have been developed to detect specific types of sRNA loci in the genome and quantify their differential expression. While microRNAs (miRNAs) are not known to play a large role in transposon regulation, a large fraction of miRNAs and other known TE regulatory sRNAs are present in multiple copies in the genome, making TE-focused strategies for multimapped read resolution useful, even for non TE-derived sRNAs. Statistical techniques, including machine learning, have already been extensively employed in the arena of piRNA prediction, a critical step for the ultimate quantification of piRNA reads accumulation in packages like piRNAPredictor (Li et al. 2016), Piano (Wang et al.

2014), and a k-mer based method described by Zhang et al. (Zhang et al. 2011) ShortStack after publication was updated to include Butter (Axtell 2014) which now performs statistical redistribution of multimapped reads.

These methods described above have largely considered sRNA classes separately, however several packages including Unitas (Gebert et al. 2017) and TEsmall (O'Neill et al. 2018) have strived to consider sRNA classes comprehensively to facilitate proper normalization of heterogeneous sRNA libraries, and to facilitate differential expression analysis across classes while taking into consideration ambiguously mapped reads.

While several iterative statistical methods have been employed in the study of sRNAs for annotation and target prediction (Hadi et al. 2018), there is still much room for improvement in the handling of ambiguously mapped reads for small RNA expression analysis. Many of these issues have been nicely reviewed by Bousios et al. (Bousios et al. 2017) particularly in the context of plants whose genomes are highly enriched in TEs and where sRNAs form a large component of the TE silencing machinery. Briefly, the chief challenge for applying probabilistic read redistribution algorithms for sRNA loci is that many types of sRNAs accumulate as very short transcripts cut from larger precursors. Often the precursors are rapidly processed and/or would not be caught by sRNA library preparation protocols. For miRNAs, for example, typically only the guide and passenger strands are detected in sRNA-seq libraries, leaving only two short ~22 nucleotide sRNAs and few surrounding reads from the precursor transcript to help guide decisions about the true originating locus. Thus, some loci may be more amenable to statistical inference algorithms, while others need additional assays in order to determine the precise source of sRNA biogenesis.

IP-seq (ChIP, CLIP, and RIP)

In this section, we have grouped together multiple disparate genomics data types that all involve immunoprecipitation-based steps in order to find protein binding sites in nucleic acids. These data can be derived from chromatin bound factors (ChIP-seq) or RNA-binding proteins (CLIP-seq/RIP-seq), but are grouped here as IP-seq because of the similar challenges these data types present for computational analysis pipelines. Typically, the published pipelines for IP-seq data analysis begin by discarding multimapped reads in order to achieve higher specificity and resolution for the protein binding sites. This can be troublesome when studying proteins which bind to regions rich in repetitive elements. For example, H3K9me3 histone markers are known to be enriched in constitutive heterochromatin (Zhang et al. 2015), a region of the genome highly enriched in repeat elements. Therefore, when calling H3K9me3 peaks using only uniquely mapped reads, the actual enrichment above background levels may be significantly higher than what is reported, skewing the estimates of background levels and discarding many truly bound regions. While this is a known issue for heterochromatin binding proteins, recent surveys of DNA- and RNA- factors have shown that transposon-derived regulatory elements form a significant fraction of both transcription factor binding sites (Cao et al. 2019; Sundaram et al. 2014) as well as RNA-binding protein recognition elements (Attig et al. 2018; Kelley et al. 2014).

For ChIP-seq based datasets, it is important to acknowledge the differences and difficulties associated with attempting to detect binding elements for chromatin binding factors and marked histones that typically bind broadly over large areas (broad peaks) as compared to transcription factors, which typically display sharp, narrow peaks. H3K9me3 typically shows a broad peak profile, as these histone marks are found on nucleosomes spread across wide stretches of chromatin. This distribution warrants a different detection strategy than that used for a typical transcription factor, such as MYC, which might occupy narrow binding regions, on the order of ~50-150 nucleotides in a typical assay. This is particularly relevant when these different peaks occur in repetitive genomic regions. The larger the bound region, the more likely it is that some of that genomic sequence will be uniquely mappable, which can guide the inference about read accumulation in adjacent sequences.

To address multimapped reads specifically, packages like the peak caller CSEM (Chung et al. 2011) have used expectation maximization to redistribute ambiguously mapped ChIP-seq reads based on the distribution of surrounding uniquely mapped reads. Due to the reliance on uniquely mappable reads, these methods function best on broader peaks because they query a larger region, which may be more likely to contain uniquely mappable content. LONUT (Wang et al. 2013) calls a set of unique peaks and a set of non-unique peaks, then aggregates both call sets together to remove any redundancy. MOSAiCS (Sun et al. 2013), while not specifically developed to handle repetitive regions, recommends using the CSEM algorithm as a pre-processing step in order to include multimapped reads. DROMPA (Nakato et al. 2013) and Crunch (Berger et al. 2019) take into account multimapped reads using a simple 1/n fractional distribution strategy. Crunch subsequently places a large emphasis on motif prediction and annotation. The analysis pipeline MapRRcon (Sun et al. 2018) uses unique and multimapped reads, but resolves the issue of multimapped read ambiguity by calling peaks on the consensus sequence of transposon subfamilies.

There is still significant room for progress in the arena of ChIP-seq analysis in repetitive regions. It is still difficult to call narrow peaks in repetitive regions, due to the lack of sufficient reads surrounding the locus of interest to guide the inference algorithms. Perm-seq(Zeng et al. 2015) addresses this issue by using the orthogonal dataset of DNAase hypersensitivity profiling for better resolution in repetitive regions of the genome. As sufficient reference datasets become available in multiple cell types and conditions, this may make this strategy feasible as a general method. In contrast, while broad peak callers tend to include more information within the locus of interest to help guide inference across repetitive regions, the data from these methods tend to have a lower signal-to-noise ratio, such that improvement of broad peak callers generally is still an active area of computational development.

The problems described above in the context of ChIP-seq analysis are compounded in the context of CLIP- and RIP-seq datasets, where one must also normalize for differences in the expression level of the bound transcript substrates. If the bound transcripts contain repetitive regions, or are entirely composed of repetitive elements, one must first find a way to accurately distribute ambiguous reads among the input transcriptome dataset before calling enriched binding sites in particular transcripts. CLIPper (Lovci et al. 2013) was one of the first CLIP-seq pipelines, but was restricted to uniquely mapped reads only. CLIPSeqTools (Maragkakis et al. 2016) is a CLIP- analysis pipeline which randomly assigns ambiguously mapped reads to one of their candidate mapping loci. CLAM (Zhang and Xing 2017) uses expectation maximization algorithms, as described above, to redistribute ambiguously mapped reads between expressed transcripts, but the algorithm works only on the alignment file and does not include information about enriched peaks in its statistical weights. PROBer (Li et al. 2017) has been developed as a general purpose algorithm for detecting sites of RNA binding or modification (termed 'toeprint' profiling) and includes an algorithm for handling multimapped reads using a Gibbs sampler approach to iteratively infer a single “best” alignment for each read. While PROBer does include steps to handle multimapped reads, it was not developed specifically for TEs, and thus has not been tested on highly repetitive regions, such as TEs that are very close to the consensus.

DNA methylation-seq

We have detailed several methods to asses differential expression, and protein binding in the context of repetitive elements. However, a critical component to the understanding of transposon biology is the analysis of DNA methylation as it is the main mechanism by which transposons are transcriptionally silenced long term (Deniz et al. 2019). To assess DNA methylation, particularly the 5-methylcytosine (5-mC) modification, several techniques have been developed and compared (Harris et al. 2010). In brief, the most common method to assess DNA methylation is bisulfite sequencing: whole genome DNA sequencing following bisulfite conversion of all non-methylated cytosine residues to uracil. Bisulfite sequencing based methods can be non-directional (Cokus et al. 2008), or directional (Lister et al. 2008) allowing one to reduce the ambiguity of strand of origin. One of the first analysis pipelines developed for high-throughput bisulfite sequencing was in *Arabidopsis* (Lister et al. 2008) and analysis was performed in conjunction with sRNA-seq datasets. In this pipeline, ambiguously mapped reads were discarded by mapping to a repeat masked version of the genome, a technique once commonly used in animal systems to reduce mapping ambiguity in the context of bisulfite induced C>T conversions (Lister et al. 2009). Bisulfite sequencing analysis differs significantly from other analysis pipelines in that often two reference genomes are used, one which contains converted cytosines in addition to the original reference genome. In this context, what are considered ambiguous reads are those reads which map to both the converted and unconverted reference genomes. This compounds the difficulty of assigning multimapped reads, such that many published bisulfite sequencing software packages choose not to include multimapped reads to avoid this confounded ambiguity (see Table 1). The most commonly used pipelines for bisulfite sequencing reads including BSMAP (Xi and Li 2009), Bismark (Krueger and Andrews 2011), MOABS (Sun et al. 2014), and BS-Seeker3 (Huang et al. 2018), none of which include probabilistic handling of multimapper reads. For a more comprehensive list of non-TE-specific methylation pipelines, please see the review by Adusumalli et al. (Adusumalli et al. 2014) and the supplemental material of a recently published pipeline, bicycle (Graña et al. 2018). Here, confounding between ambiguity in bisulfite conversion rates, non-reference polymorphisms, and read non-uniqueness can complicate the statistical tests used to determine if a site in the genome is differentially methylated. Thus, this represents an area of computational genomics that could benefit greatly from further development.

Since DNA methylation is a critical mechanism by which transposons are silenced, several groups have used new methods to improve methylation analysis for TEs. TEPID (Stuart et al. 2016) and epiTEome (Daron and Slotkin 2017) were designed to improve analysis of TE methylation levels by including the analysis of split reads that cross junctions between TEs and uniquely mappable genome regions. An approach employed to assess the low mappability of young transposable elements, like L1-Ta, in the human genome was repurposed to align bisulfite reads to a consensus sequence as described in Shukla et al. (Shukla et al. 2013). One interesting method to improve methylation analysis is to first rigorously determine the average bisulfite conversion rates genome-wide, then use this as a parameter to tease apart mapping ambiguities from differences in conversion rates, as done by Noshay et al. (Noshay et al. 2019). Despite these improvements, DNA methylation analysis is still a difficult bioinformatic challenge that would benefit from further study.

Single Cell RNA-seq

All of the software described above has been geared towards genomics datasets generated from bulk tissue samples. However, bulk profiling of heterogeneous cell populations only provides averages that obscure underlying variability of TE expression across cell types, as illustrated in Figure 3. This problem is further amplified when aggregating transcriptional signal across numerous loci within high copy-number TE families. It remains largely unknown how TE de-repression varies between individual cells, what factors drive such differences, and how this variability might affect cellular phenotypes. Single cell RNA-sequencing (scRNA-seq) promises to answer some of those questions and has already redefined our knowledge about tissue composition and gene regulatory networks (Tanay and Regev 2017). While its broad application has so far been largely restricted to the study of gene activity patterns, a few pioneering studies have utilized first-generation protocols to identify TE expression dynamics across single pre-implantation embryonic cells (Göke et al. 2015; Boroviak et al. 2018). Those early efforts were largely limited by small cell numbers, high sequencing burden per cell, and lack of molecular barcode counts to estimate true transcriptional output, thus preventing broad-scale adaptation. Since then, the increasing demand in single cell transcriptome data has seen an unprecedented expansion of available scRNA-seq protocols with considerably improved throughput, robustness, and error-rates (Ziegenhain et al. 2017). One such publication was by Guo et al. (Guo et al. 2018) where the number of cells were scaled up allowing for investigation into TE dynamics in spermatogenesis.

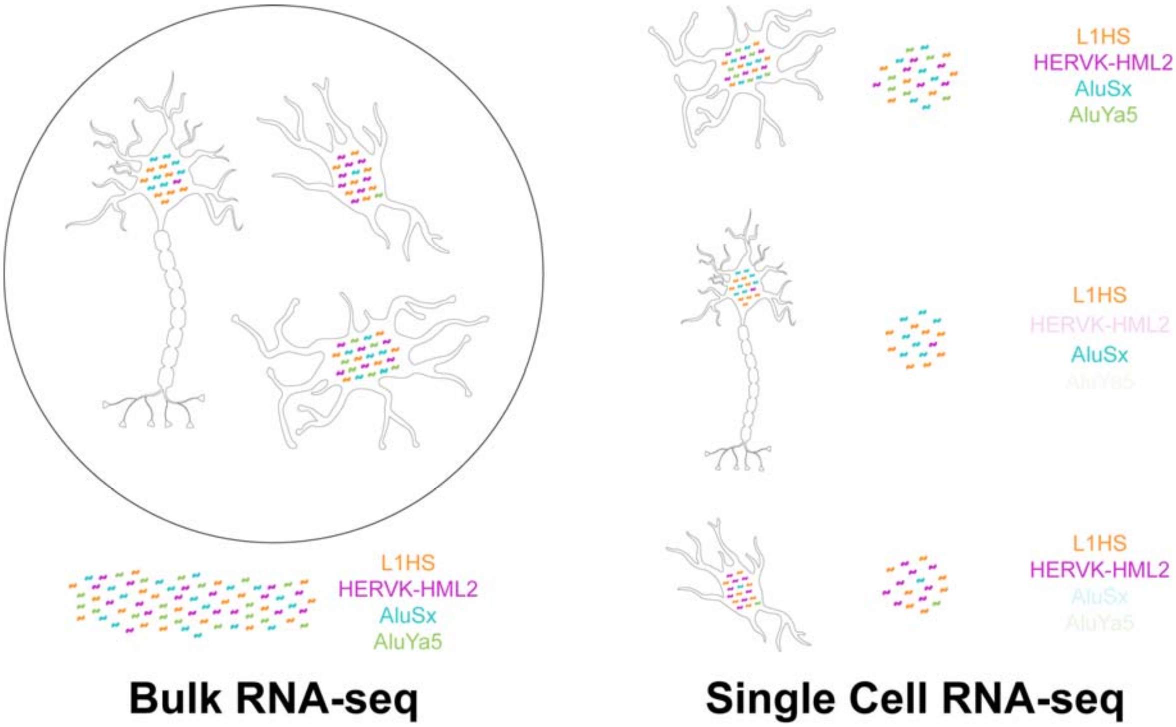


Figure 3 Comparison of bulk RNA-seq versus single cell RNA-seq

Heterogeneity in expression profiles across cell types is masked by bulk sequencing methods. Transposable element (TE) expression may vary across cell types, between cells of the same type, and within the same cells across time. Single cell methods are necessary to reveal this heterogeneity, but software for single cell data analysis is not currently optimized for handling TEs.

Despite such experimental advancements, inherent design principles of scRNA-seq protocols that cooperate with the well-known challenges of TE transcriptome analysis have so far prevented their common application for the study of TE expression at single cell resolution (Fig 4). For example, many popular methods quantify RNA molecules at the 3′ end of polyadenylated mRNAs (Jaitin et al. 2014; Macosko et al. 2015; Hashimshony et al. 2016; Nakamura et al. 2015; Cao et al. 2017) and therefore depend on accurate reference models to bridge the gap between polyadenylation sites and the corresponding transcript isoform and/or promoter. This is problematic for TE-derived transcripts, which are generally poorly annotated in many species. While protocols with full-length transcript coverage might alleviate some of those problems, the naïve assignment of reads to the nearest TE interval can still lead to erroneous assignment, misattribution of intronic reads from unprocessed pre-mRNAs, and hence misinterpretation of TE de-repression. Full-length protocols additionally suffer from higher sequencing burden, often lack of unique molecular identifiers to account for PCR duplicates, and potentially higher background TE read coverage due to intronic signal originating from pre-mRNAs (Manno et al. 2018; Ding et al. 2019).

A potential solution to minimize misattribution problems are 5′ end based scRNA-seq protocols that incorporate a template switch oligo (TSO) towards the start of transcription initiation (Islam et al. 2012; Cole et al. 2018). Although incomplete processing and premature TSO incorporation during library preparation might vary between transcripts and cells, such protocols have already been successfully used to map alternative transcription start sites between individual cells (Karlsson et al. 2017). Importantly, a recent study also demonstrated its utility to quantify unexpected variability in TE promoter activity between thousands of single cancer cells following epigenetic therapy(Brocks et al.

2018). However, the problem of premature TSO incorporation, combined with the pervasive nature of TEs, and technical noise inherent to all current scRNA-seq protocols requires dedicated strategies to mitigate the danger of spurious estimates of TE cell-to-cell variation. To the best of our knowledge, no peer-reviewed computational pipeline currently combines such features with the reliable quantification of TEs at single cell resolution, but unpublished efforts already aim to facilitate TE single cell analysis for a wide array of available scRNA-seq protocols (https://tanaylab.github.io/Repsc/). With the continuous methodological advancements and the increasing interest in TE biology, we anticipate a rapid progress towards the routine quantification of TEs in individual cells that will be accompanied by the discovery of unprecedented heterogeneity in TE transcription patterns.

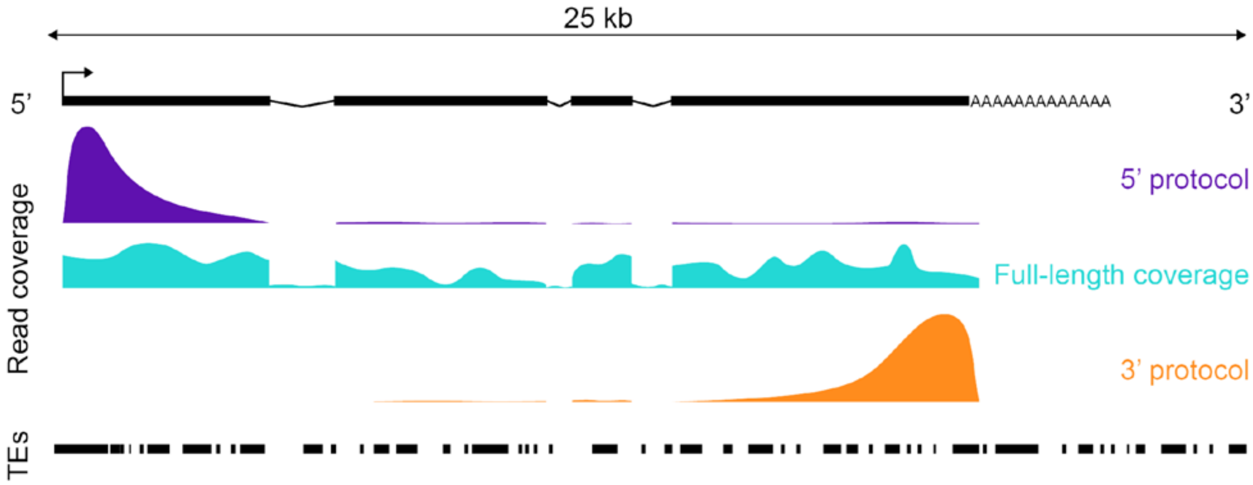


Figure 4 Impact of different RNA-seq library strategies on read coverage along a TE-derived transcript with exon/intron structure

(top). TE intervals are shown at the bottom. Briefly, scRNA-seq protocols that offer full-length transcript coverage provide the best means to identify full length transcribed TEs in a locus-specific manner, but this method suffers from noise due to intronic TEs in host genes that might be mistaken for expressed TE transcripts as well as the inability to barcode individual mRNA molecules. 5′ and 3′ based protocols allow for barcodes that enable mRNA molecule counting, with 5′ protocols also offering the ability to detect TE transcripts originating from proper TE promoters.

Conclusions

What is now doable?

The last years have seen a general improvement in sequencing read length, making it possible to study the majority of transposable elements in a genome wide fashion. For particularly young and less diverged families, we have discussed at length the strides made in genome biology to address the difficulties of treating ambiguously mapped sequencing fragments for differential expression and binding analyses. In the context of highly repetitive regions of the genome these difficulties are compounded, particularly for the most active TEs, which remain close to their consensus sequence and thus are the most difficult to map. The greatest progress has been made with RNA-seq data analysis, as we have progressed from using simple fractional assignments of multimapped reads within genes to approaching true locus specific resolution in the most repetitive regions of the genome – such as the L1HS subfamily, active Alu families, and composite SVA elements. Progress has been made in the realm of sRNA analysis as these improved algorithms for RNA-seq analysis have now been incorporated into sRNA-seq data analysis pipelines. In immunoprecipitation based assays, for ChIP- and CLIP-seq datasets, efforts have been made to use probabilistic read redistribution for peaks within repetitive regions, but challenges remain.

What is still hard?

sRNA-seq data contains a large proportion of multimapped reads, and while significant effort has been put forth to leverage advanced iterative statistical methods for novel sRNA discovery and target prediction, these methods have not been as widely applied to sRNA-seq transcript quantification. This may be attributed to the tight distribution of sRNA reads across their mapping loci, making it difficult to garner locus-specific information from adjacent reads. Moreover, these much shorter reads (18-30 nt) are intrinsically less unique in the genome than longer sequences.

In ChIP-seq data, the expected profile of read distributions can vary widely from the typically tall, narrow peaks associated with most transcription factor profiles or RNA-binding proteins to the broader, shorter, and noisier peaks associated with some marked histones, such as H3K9me3. Algorithms have been developed to address both types of ChIP-seq profiles. Yet the lines between these categories can be blurred, and there is a large tradeoff between the window size in peak calling and the ability to use uniquely mapped reads to probabilistically reassign all other reads to a particular locus. One area of active research for broader regions would be to incorporate multimapped reads into segmentation models which allow for the detection of changes in peak landscape, as opposed to simply calling the absence or presence of individual peaks.

Single cell RNA-seq (scRNA-seq) represents one of the newest genomic assays to be used for TE expression profiling, and as such, remains an area of greatest need for improvements in software packages specifically designed to handle the complexities inherent in TE genomics. Efforts are already underway, but as yet no published software packages for scRNA-seq are available. That said, many standard scRNA-seq packages could be adapted for this use, as in the example protocol described above. However, as discussed in detail, differences in the experimental protocols used to generate scRNA-seq libraries will have a large impact upon the interpretability of the data, and this is particularly problematic for TE expression analysis.

Two types of analysis which largely do not include multimapped reads are assays for transposase accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al. 2015) and Hi-C(Lieberman-Aiden et al. 2009), an extension of chromosome conformation capture (3C). The read distributions for ATAC-seq data greatly resemble those of ChIP-seq and this analysis encounters similar computational difficulties when studying repetitive regions of the genome. Fortunately, as this analysis is similar to ChIP-seq there has already been significant effort which could be incorporated into ATAC-seq analysis. Adapting Hi-C pipelines to take into account multimapped reads is still a difficult task as this type of analysis already requires the resolution of chimeric reads representing genomic proximity. mHiC (Zheng et al. 2019) has been developed to address this issue, but the relative sensitivity to highly repetitive transposon regions is unclear. Significant work has been done using these methods to address the role of transposons in genome architecture and the transition from the embryonic cell state to early embryonic like cells(Cao et al. 2018; Kruse et al. 2019; Rodriguez-Terrones et al. 2018). These analyses can only improve as better methods for handling repetitive reads are included.

What new technology needs to be developed?

Long read sequencing technologies promise to solve many issues inherent in the assays described above. Once the issues with throughput and error rates can be solved, long read sequencing would enable the isolation of entire transcripts and, if correctly barcoded, would also allow for accurately calibrated expression estimates. These technologies could also be combined with antibody based pulldowns and endonuclease based footprinting assays, to accurately call cis-regulatory regions derived from TEs. Finally, long-read genome resequencing assays that sequence through highly repetitive genome regions may allow for better genomic annotations that will benefit all of the applications described above. To this end, not only must new experimental protocols be developed which emphasize longer reads, but new computational pipelines must also be developed to ensure that these long read analysis pipelines properly handle and account for the complications inherent in addressing TE genomics.

Acknowledgments

We would like to acknowledge Ying Jin for her helpful conversations about TE genomics. We apologize for any software we failed to cite in this review pertinent to the study of repetitive elements in genome biology. Schematic images were adapted with permission from Servier Medical Art (https://smart.servier.com). MGH receives support from the Rita Allen Foundation and the Chan-Zuckerberg Initiative. KO acknowledges funding from the National Science Foundation Graduate Research Fellowship and NIH training grant (2T32GM065094-16). DB is supported by an EMBO long-term fellowship (ALTF 900-2017).

Competing Interests

The authors have no competing interests to declare.

Author Contributions

All authors contributed to the writing and editing of this manuscript.

2.2.2 Supplemental Table

Supplemental table is available at

https://royalsocietypublishing.org/doi/suppl/10.1098/rstb.2019.0345.

2.3 Conclusion

This review provides an overview of the state of computational methods for transposon-aware analysis of RNA-seq, small-RNA-seq, ChIP-seq, Clip-seq and DNA methylome-seq data as of October 2020. In the next chapter, I will be discussing in greater detail a subset of this review which was dedicated to small RNA-seq library analysis. This will include the paper associated with TEsmall, an analysis pipeline briefly mentioned in this review, improvements to the TEsmall pipeline, and preliminary analysis into the effect of TDP-43 loss of function in adult human somatic cells on small RNA biology.

Chapter 3: TEsmall and description of TDP-43 LOF dependent small RNA biology in human adult somatic tissue

3.1 An introduction to TEsmall

This chapter begins with a reformatted version of the paper, published in October of 2018 in Frontiers in Genetics, describing the TEsmall analysis pipeline. I was co-first author of the of the publication with Wen-Wei Liao, a previous member of our lab, and was responsible for composing the article, the generation of figures, and addressing revisions. I additionally contributed to code development and packaging for the TEsmall pipeline. Following the article, I discuss improvements to the pipeline and analysis performed with the original TEsmall pipeline investigating TDP-43 loss of function associated biology in an adult human somatic cell line.

3.2.1 The TEsmall manuscript

TEsmall identifies small RNAs associated with targeted inhibitor resistance in melanoma

Kathryn O'Neill*, Wen-Wei Liao*, Ami Patel and Molly Gale Hammell

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

*These authors contributed equally to the work

MicroRNAs (miRNAs) are small 21-22nt RNAs that act to regulate the expression of mRNA target genes through direct binding to mRNA targets. While miRNAs typically dominate small RNA transcriptomes, many other classes are present including tRNAs,

snoRNAs, snRNAs, Y-RNAs, piRNAs, and siRNAs. Interactions between processing machinery and targeting networks of these various small RNA classes remains unclear, largely because these small RNAs are typically analyzed separately. Here we present TEsmall, a tool that allows for the simultaneous processing and analysis of small RNAs (sRNAs) from each annotated class in a single integrated workflow. The pipeline begins with raw fastq reads and proceeds all the way to producing count tables formatted for differential expression analysis. Several interactive charts are also produced to look at overall distributions in length and annotation classes. We next applied the TEsmall pipeline to small RNA libraries generated from melanoma cells responding to targeted inhibitors of the MAPK pathway. Targeted oncogene inhibitors have emerged as way to tailor cancer therapies to the particular mutations present in a given tumor. While these targeted strategies are typically effective for short intervals, the emergence of resistance is extremely common, limiting the effectiveness of single-agent therapeutics and driving the need for a better understanding of resistance mechanisms. Using TEsmall, we identified several microRNAs and other small RNA classes that are enriched in inhibitor resistant melanoma cells in multiple melanoma cell lines and may be able to serve as markers of resistant populations more generally.

Introduction

microRNAs (miRNAs) are small 21-22 nucleotide RNA molecules which have been shown to play a critical role in metazoan development and gene regulation. While typically derived from short hairpin RNA precursors located in both intergenic and intronic regions, miRNAs can also be processed from other ncRNAs including tRNAs and spliced intron lariats.(Bartel 2018; Schorn et al. 2017) In addition to governing development, small RNAs (sRNAs) play a critical role in repressing transcripts derived from repetitive regions of the genome. In animals, siRNAs and piRNAs function to repress transposons in somatic cells, and the germline respectively (Ghildiyal et al. 2008; Castañeda et al. 2011). Identification of miRNAs and siRNAs which originate from non-canonical regions of the genome is more challenging with few programs designed to detect sRNAs from all classes in both unique and repetitive genomic loci. It is for this reason we present TEsmall, a package designed specifically for the simultaneous analysis of sRNAs derived from a variety of genomic features. In particular, this package facilitates the discovery of intriguing biological phenomena otherwise masked by insufficient annotation of repetitive genomic elements, such as siRNAs, and allows these elements to be easily incorporated into downstream differential analysis through packages like DESeq2 (Love et al. 2014).

We have tested the ability of TEsmall to characterize the expression profiles of small RNAs from a variety of classes in the context of melanoma cell lines responding to targeted inhibitors of the BRAF oncogene. The genetic basis of melanoma development is fairly well understood, with activating mutations in the oncogene BRAF occurring in a majority of melanoma patient tumors (Hodis et al. 2012), which also harbor hundreds of secondary mutations of unknown impact. Specific inhibitors that target activated BRAF as well as the downstream MAPK/ERK signaling pathway have been developed, which dramatically reduce the growth of melanoma cells in patients. However, the effects of these drugs typically extend patient lifespan for six months or less, as the tumors rapidly develop resistance to these targeted therapies (Villanueva et al. 2010). While some tumors resistant to BRAF inhibitors acquire additional genetic lesions that elevate MAPK or AKT signaling (Alcala and Flaherty 2012), many therapy-resistant cell lines establish resistance without a clearly understood mechanism of resistance (Gatenby and Brown 2018). Changes to small RNA profiles in melanoma cells responding to targeted inhibitors is an especially poorly understood subset of the genomic and transcriptomic changes that occur. To understand how small RNA alterations might contribute to the development of resistance to BRAF inhibitors in 451Lu melanoma cells that carry BRAFV600E mutations, we undertook a small RNA sequencing study of cells before and after the establishment of BRAF inhibitor resistance.

Materials and Methods

Melanoma cell culture

In this dataset, 451 Lu patient derived melanoma cell lines were used to explore the sRNA profiles of cells that are either sensitive or resistant to small molecule inhibitors of the BRAF kinase. Specifically, the melanoma patient derived 451Lu-Par cells are grown in standard growth media (DMEM with 10% FBS), while the 451Lu-BR cells are grown in standard growth medium supplemented with a 1uM concentration of the BRAF inhibitor vemurafenib. Both cell lines are adherent cells grown in standard 2D cell culture. The derivation of BRAF inhibitor resistance in these cells lines is described by Villanueva et al. (Villanueva et al. 2010) and the cell lines are available from Rockland for both 451Lu cells (cat: 451Lu-01-0001) and 451Lu-BR cells (cat: 451Lu BR-01-0001).

Small RNA sequencing libraries

Total RNA was extracted using the Ambion PureLink RNA Mini Kit to extract up to 2 μg of total RNA from ~1x10^6 melanoma cells from either the 451Lu-Par or 451Lu-BR lines. Following Bioanalyzer verification of RIN numbers at or above ~9, the RNA extracts were next used to create small RNA sequencing libraries. The small RNA sequencing libraries were prepared with the Illumina TruSeq Small RNA Library Preparation Kits using an input of 1.2

µg total RNA and following the manufacturer's protocols as described, using 15 PCR cycles to reduce the likelihood of PCR amplification artifacts. The libraries were pooled and indexed with 6nt Illumina barcodes, such that 6 libraries could be sequenced per lane on an Illumina Genome Analyzer IIx. The reads were sequenced as single-end 50bp reads, to a depth of approximately 35 million reads per library. The dataset is available through GEO at the following accession number: GSE116134. A table of sequenced and mapped read counts for each library is presented in Supplemental File 1.

qPCR Validation

Taqman qPCR assays were used to validate the analysis results of TEsmall for a subset of microRNAs. Specifically, standard Taqman qPCR probes were obtained for the following microRNAs: miR-100, miR-184, and miR-211. Control probes were obtained for RNU58 and the U6 small RNA. Custom Taqman probes were obtained for the predicted mature sequence of the novel candidate miRtron derived from the VIM intron 6 locus. The Taqman protocol was followed as described in the Thermo Fisher Scientific TaqMan MicroRNA Assay protocol, available from the manufacturer. Briefly, 10 ng of total RNA was used as the input to a microRNA specific reverse transcription (RT) assay using the TaqMan MicroRNA Reverse Transcription Kit and an RT primer specific to the miRNA of interest. Next, qPCR amplification was performed using 1.33 uL of the output from the RT reaction, TaqMan Universal PCR Master Mix II no UNG, and the TaqMan Small RNA Assay Kit specific to the miRNA of interest. Each small RNA specific assay was performed with input RNA from two biological replicates of the 451Lu-Par and 451Lu-BR cell lines, with 3 technical replicates per biological replicate, on a Thermo Fisher StepOnePlus Real-Time PCR System. The "Comparative Ct" analysis method described in the manufacturer's protocol was used for calculating fold change, standard deviation, and t-test based P-values. Briefly, the three technical replicates for each probe were combined to create a mean Ct value per probe per sample. The average of the Ct values from the two control probes in each sample was then subtracted from each microRNA Ct value to create a normalized "Δ-Ct" value for each microRNA in each sample. Following averaging of Δ-Ct values between the two biological replicates in each condition, the Δ-Δ-Ct value was calculated as the difference in mean Δ-Ct values for the same microRNA across conditions. Fold change represents $2^{\Delta\text{-}\Delta\text{-Ct}}$, with errors on each Ct value combined quadratically.

TEsmall module

TEsmall functions by accepting raw input in FASTQ file format from next generation sequencing platforms in conjunction with genomic annotation sets via an online server. Adapters associated with siRNA library preparation are trimmed by TEsmall through the cutadapt package(Martin 2011). In order to remove degradation products from abundant ribosomal RNAs, rRNA derived reads are next filtered from the data before proceeding to analysis. This mapping step allows for up to 2 mismatches and filters a single alignment per read specified by the option: bowtie -v 2 -k 1 using bowtie (v1.2.1)(Langmead 2010). Small RNA reads remaining after rRNA filtering are then aligned more stringently, disallowing mismatches, option: bowtie -v 0 -a -m 100. All alignments in this step which map to fewer than 100 genomic loci are reported allowing for the classification of multimapper reads common to sRNA data, in particular structural RNAs like tRNAs and transposable element targeting siRNAs. Following alignment to the genome, each alignment is annotated via a sequential decision tree, as follows. The alignments are distributed to each annotation category in order, then removed from the pool of alignments in order to facilitate priority annotation of, for example, intronic microRNA reads that should properly be annotated as microRNAs rather than intronic RNAs. The default order is: structural RNAs, miRNAs and hairpins, exons, sense transposons, antisense transposons, introns, and ultimately annotated piRNA clusters. This process is depicted in Supplemental Figure 1. This annotation class priority can be re-ordered by the user to suit the application and user preferences. An HTML output file is then created using python based Bokeh tools(Bokeh Development Team 2014) to visualize the abundance distributions, length distributions, and mapping logs of all small RNAs in the dataset (Fig. 5). In conjunction with this HTML output, TEsmall compiles multiple flat text output files, including a counts file that is structured to be directly compatible with DESeq2 (Love et al. 2014) for differential analysis. The abundance calculations for these count files are 1/n normalized at the end of this annotation process, where n represents the number of alignments per read, to ensure no double-counting of multimappers. Additional packages employed within the TEsmall workflow include bedtools (Quinlan and Hall 2010), pandas (McKinney 2010), samtools (Li et al. 2009), pybedtools (Dale et al. 2011), and scipy (Jones et al.).

The TEsmall code is available open-source from GitHub at the following location: https://github.com/mhammell-laboratory/tesmall.

Annotation files for the human (hg19) and fly (dm3) genomes can be found at the following location: http://labshare.cshl.edu/shares/mhammelllab/www-data/TEsmall/.

Instructions for creating annotation files for additional genomes can be found in Supplemental File 4.

Differential Analysis with DESeq2

The counts file produced by TEsmall were subsequently imported into DESeq2 (v1.18.1) to perform differential analysis between 451Lu-PAR and 451Lu-BR cell lines, as follows. The counts file was filtered to remove low abundance species (< 20 counts across all libraries) and increase the sensitivity of DESeq2. Normalization of the counts for differential analysis was performed using the default DESeq2 method during statistical analysis. For downstream visualization, the counts were normalized by the built-in variance stabilizing transformation (VST) method in DESeq2. Small RNAs with an adjusted P-value < 0.05 were considered statistically significant. The full DESeq2 output file is given in Supplemental File 2.

Visualization

Figures were produced using the R packages ggplot2 (Wickham 2009), gplots (Warnes et al.) and GenomicRanges (Lawrence et al. 2013) for scatterplots, heatmaps, and wiggleplots respectively. Python package matplotlib (Hunter 2007) was used for all barplots. RNA secondary structures were rendered using the forna webtool (Kerpedjiev et al. 2015), secondary structure for the Arg-ACG-1-2 tRNA was pulled from the UCSC GtRNAdb tRNA covariance model, and structure of vimentin intron 6 was predicted using RNAfold's minimum free energy model. (Gruber et al. 2008)

Results

TEsmall Workflow

TEsmall is a package specifically designed to identify sRNAs derived from a variety of genomic features simultaneously, such that users can evaluate the relative abundances and profiles of many sources of small RNAs on a common scale in a single pipeline. This serves as a novel improvement to currently available software such as mirDeep2 (Friedländer et al. 2012) and piPipes (Han et al. 2015), which are optimized for the analysis of miRNAs and piRNAs respectively, but are not equipped to evaluate both types of small RNAs together. TEsmall is also designed so that its output is optimally formatted for downstream differential analysis with statistical modeling software, such as DEseq2 (Anders and Huber 2010). A flowchart describing the entire TEsmall workflow is given in Figure 5A, with example output charts given in panels 1B-E. Specifically, in the first module of TEsmall, raw small RNA sequencing reads from Illumina NGS sequencing platforms serve as the input without the need to pre-process the data before beginning analysis. TEsmall first trims adapter contaminants from the reads and then filters the reads for appropriate size ranges, with a default of 16-36 nucleotides in length. The next module of TEsmall removes contaminating ribosomal RNA fragments by mapping with bowtie (Langmead 2010) to a library of rRNAs for the specified genome. Removal of rRNA reads is critical as rRNA degradation products are a major source of contamination for sRNA data. Remaining reads are mapped to the genome, with a default mapping strategy optimized for repetitive regions with up to 100 alignments per read, though this may be altered by the user. The reads are next sequentially annotated to several small RNA classes and genomic features, with a decision tree implemented to prioritize annotation categories. This has the goal of attributing reads mapping to intronic microRNAs as “microRNA” reads, for example, rather than annotating these reads as having an intronic source. Following annotation of each read, aggregate abundances are calculated for each sequencing library and outputted as a counts table suitable for downstream differential expression analysis. Importantly, any multimapper reads in these counts tables are weighted according to the number of genomic loci from which they derive (1/n where n is the number of alignments) to avoid any double-counting of multimapper reads in the counts tables. In addition to an output file including all raw count data per sample, RNA species ID, and type classification, TEsmall provides an aesthetic output HTML (Fig. 5B) summarizing distribution of read lengths (Fig. 5C), proportion of reads assigned to each sRNA type (Fig. 5D), and distribution of reads of a particular size to each of the sRNA categories (Fig. 5E). In addition to these summary plots, TEsmall presents a table with summary statistics of read proportion, raw input and trimmed read counts to quickly assess any potential biases in library preparation that may affect downstream normalization.

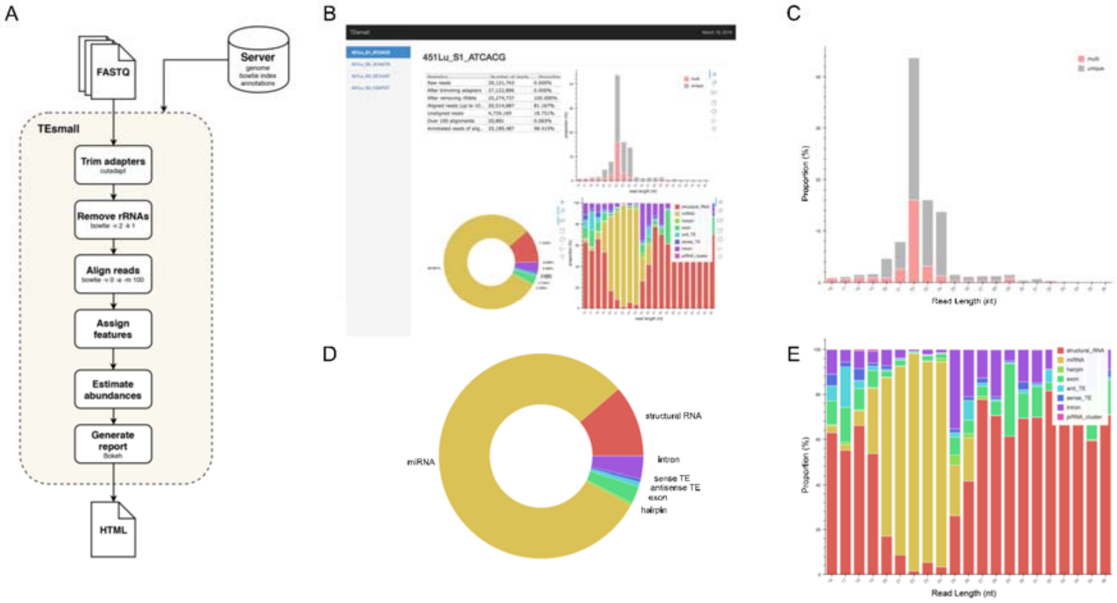


Figure 5 Flow chart and output HTML of TEsmall

(A) Flow chart of TEsmall's treatment of input high-throughput sequencing data, input genome indices, and output. (B) Screenshot of HTML output file for one sample (C) Bar plot depicting size distribution of unique and multimapper reads. (D) Circle plot depicting distribution of reads to each subtype. (E) Bar plot depicting proportion of subtypes across read length.

Application of TEsmall to melanoma sRNA profiles

As described, drug resistance is a known hurdle in the treatment of melanoma, driving the need for a better understanding of how cells develop resistance. We have chosen to investigate the alterations in sRNA profiles, as one marker of cellular state. To investigate the effect of BRAF kinase drug resistance on sRNA composition in patient derived melanoma cell lines, we performed differential expression analysis following classification by TEsmall in two biological replicates of parental and BRAF inhibitor resistant cell lines. Resistant lines were derived through exposure of 451Lu patient derived parental cell lines to increasing concentrations of vemurafenib up to 1uM. Resistant clones were selected and expanded before exposure to an increase in vemurafenib. Cells were otherwise treated as described in Villanueva et al.(Villanueva et al. 2010) Raw count data was normalized as described in Materials and Methods by DESeq2. All VST normalized counts were averaged between parental or resistant replicates and plotted against each other to visualize trends of expression across sRNA subtypes without filtering for significant or abundant transcripts, significantly differentially expressed transcripts are represented by solid coloring (Fig. 6). Overall, there appears to be a trend towards lower expression of many sRNA classes in the 451Lu BRAF resistant samples, with more down-regulated than up-regulated species for most classes of sRNAs (Fig. 6 and Fig. 7B). However, upon testing by a two tailed Welch's t-test only structural RNAs were shown to be significantly downregulated as a class in BRAF inhibitor resistant libraries ($P< 3.52x10^{-13}$). Upon filtering for the most abundant (base mean across all replicates > 500) and significantly differentially expressed transcripts (multiple-hypothesis testing adjusted p-value < .05), trends of lower sRNA expression in the BRAF resistant samples was still seen for many intronic, exonic, and transposon mapped sRNA species (Fig. 7). Interestingly, miRNAs show an even distribution of species with negative and positive log fold changes, and since miRNAs were the most abundantly sequenced small RNAs in the libraries, this rules out a normalization issue as the explanation for down-regulated small RNAs in the other classes. It may be of interest that, after filtering for significance (P<0.05) and abundances greater than 500 reads per million mapped (RPM), structural RNAs with a negative log fold change are almost exclusively tRNAs and those with a positive log fold change are almost exclusively snoRNAs. Details of the particular sRNAs differentially expressed in each of these classes are given in Supplemental File 2.

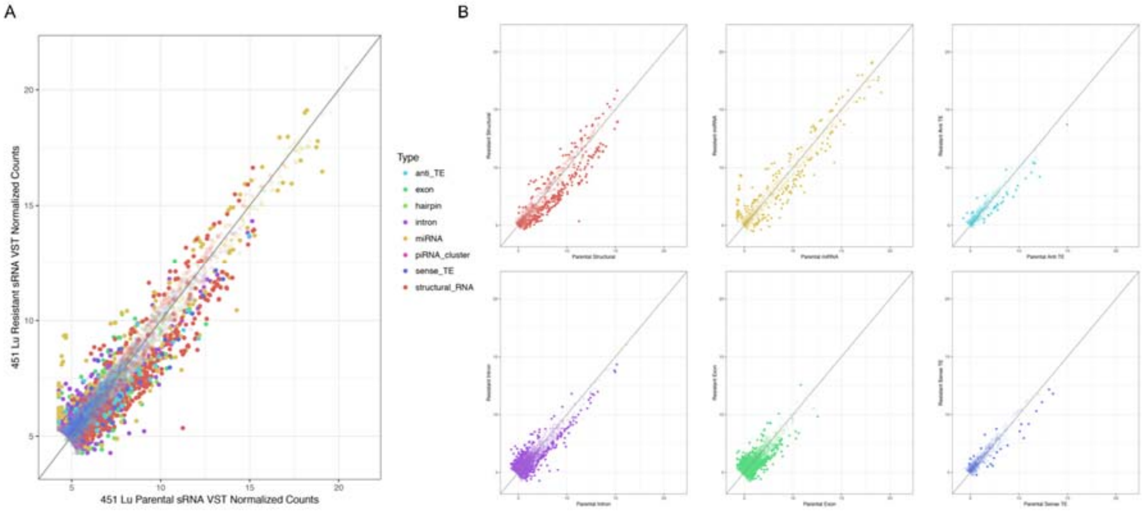


Figure 6 Scatterplots depicting 451Lu BRAF resistant versus parental RNA: VST normalized counts

(A) Overlay of all subtype scatterplots. (B) RNA subtype specific scatterplots. Transparent points represent RNA species with adjusted p-value < .05.

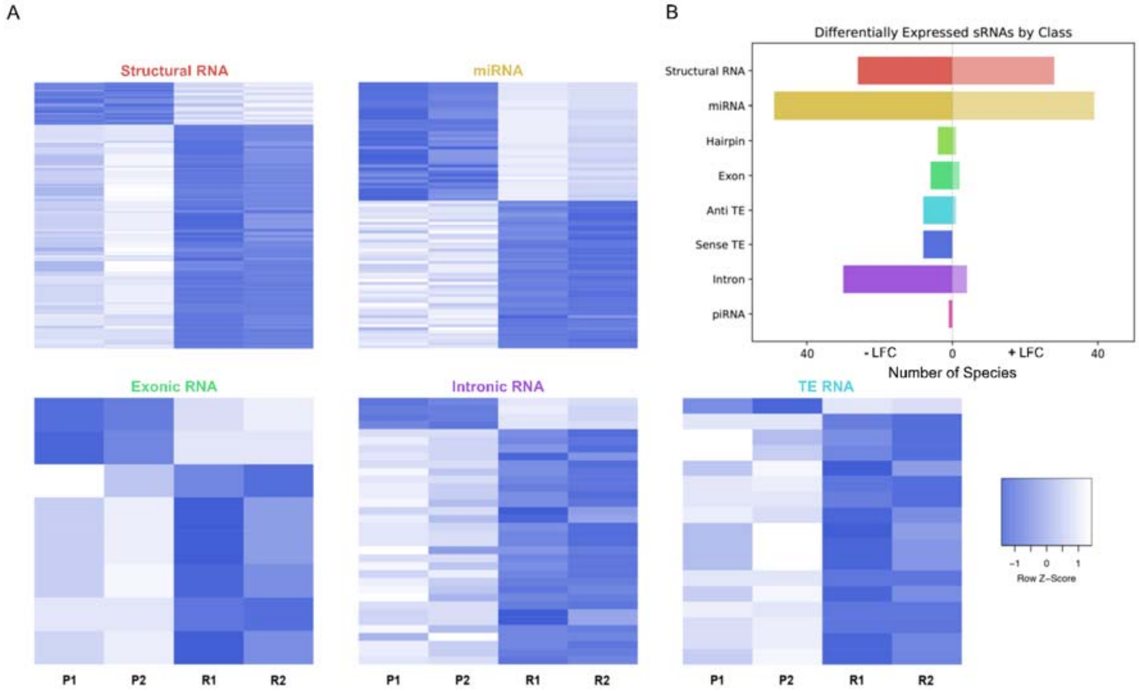


Figure 7 Differential expression analysis of highly abundant sRNAs

(A) Heatmaps depicting all significantly differentially expressed sRNAs per subtype. All heatmaps are row-scaled to lie between -1 and 1, based on the VST normalized count values. P1, P2, R1, and R2 represent 451Lu parental replicates 1 and 2 and 451Lu BRAF Resistant replicates 1 and 2, respectively. (B) Bar plot depicting number of abundant and significantly differentially expressed species with a negative or positive log fold change per subtype. Structural RNA species are collapsed for duplicate tRNAs.

Type specific analysis of sRNA species and Validation with qPCR

Several miRNAs which are significantly differentially expressed in our dataset have been previously described in the literature as playing critical roles in melanoma progression or epidermal differentiation. This includes the miRNAs miR-184, miR-211, and miR-100. In other contexts, miR-184 has been shown to arrest epidermal differentiation through de-repression of Notch in normal human keratinocytes and murine epidermis (Nagosa et al. 2017). While expression of Notch in keratinocytes is known to have a tumor suppressive phenotype, its expression has the opposite effect in melanocytes through upregulation of the PI3K/Akt and MAPK pathways (Pinnix and Herlyn 2007). Our data showed an approximate 5-fold increase in miR-184 expression in BRAF inhibitor resistant cells relative to parental (Fig. 8A and Supplemental File 2), consistent with a model where MAPK pathway activation provides a mechanism for BRAF inhibitor resistance (Villanueva et al. 2010, 2011). It has also been shown that BRAF inhibitor resistance can be mediated by regulatory escape of the transcription factor MITF from the MAPK pathway, where MITF overexpression itself conferred resistance in several melanoma cell lines (Van Allen et al. 2014). Consistent with a high MITF state, our data shows a significant upregulation of miR-211, derived from the MITF activated gene melastatin, and a significant downregulation of miR-222, known to be inversely correlated with MITF expression (Golan et al. 2015). Finally, miR-100 was also shown to be significantly downregulated in our data; this was a miRNA of interest as it has been implicated in prostate cancer as a repressor of the oncogene mTOR (Leite et al. 2013).

To validate the expression profiles from our TEsmall based differential expression analysis, we performed qPCR on several miRNAs of interest including miRNAs miR-184, miR-211, and miR-100, all of which recapitulated the trend observed in our small RNA-seq dataset (Fig. 8A). Reassuringly, the expression alterations of miR-204 and miR-211 seen in our data were also seen in an alternative melanoma derived cell line A375 in Díaz-Martínez et al. following induction of BRAF inhibitor resistance (Díaz-Martínez et al. 2018).

Upon further investigation into individual RNA species from different subtypes for follow up, we encountered an interesting and novel 21 nucleotide sRNA derived from the sixth intron in the vimentin gene (VIM) (Fig. 8B). Intronic microRNAs can either derive from their own precursor transcripts dubbed pri-miRNAs or can alternately derive from short spliced introns with internal hairpin structures dubbed miRtrons (Okamura et al. 2007). MiRtrons are typically generated via the splicing machinery from short introns (~100 nt) and subsequently processed by DICER in the cytoplasm, bypassing the canonical nuclear DROSHA processing steps. This is in contrast to intronically located miRNAs that derive from their own precursor pri-miRNAs and are dependent upon both DROSHA and DICER for processing. As visible in the minimum free energy secondary structure prediction by RNAfold, the candidate miRNA of interest is located in a stem loop structure which appears conducive to processing by DICER (Fig. 8C). The length of this mature sRNA (21nt), the short length of its host intron (350 nt), its abundance as a single RNA species, and its secondary structure within VIM intron 6 could all be consistent with a miRNA derived from either an intronic pri-miRNA or as a miRtron. This is particularly interesting as VIM is a known marker for the epithelial to mesenchymal transition and is well expressed in many cell types, but has not previously been shown to harbor a miRNA, suggesting this candidate VIM miRNA might represent a novel sRNA with particular abundance in melanoma cells.

In addition to miRNAs, TEsmall recognizes several other types of sRNAs. It has been previously reported that tRNA derived small RNA molecules (tRFs) can silence LTR retrotransposable elements through occupation of the primer binding site (PBS) as an adaptation of the role of tRNAs as retroviral primers (Schorn et al. 2017; Mak and Kleiman 1997). Through TEsmall, one is able to detect reads associated with tRNAs and transposable elements in the same pipeline facilitating observation of phenomena such as these. In our analysis, several species of sRNAs mapping antisense to transposable elements were significantly depleted in BRAF resistant cell lines compared to parental (Fig. 6). Upon further investigation, we were able to determine that a subset of these reads were tRFs derived from the Arg-CGY family of tRNAs (Fig. 9B). These candidate tRFs mapped to a subset of HERVs including HERV3, HERV30, MER51, and others (please see Supplemental File 3). This is consistent with previous literature showing HERV-R type retrotransposons are primed by Arg tRNAs (Fig. 9A) and that tRNA derived fragments can occupy retroviral primer binding sites to suppress transposon activity (Schorn et al. 2017; Mak and Kleiman 1997). It is important to note that the Arg-CGY sRNAs reported by TEsmall are consistent with the tRFs previously described in Schorn et al. as they are 18nt CCA-appended fragments originating from the 3′ T-arm of tRNAs (Schorn et al. 2017). This is shown graphically in Fig. 9, where the pileup of reads at an example HERV PBS locus can be seen in Fig. 9B, and the pileup of these same reads at the originating tRNA locus can be seen Fig. 9C-9D. In the tRNA profiles, other tRNA fragments including tRNA derived stress-induced RNAs (tiRNAs) can be seen outside of the tRF generating 3′ end, but these do not predominantly accumulate as a single abundant sRNA species.

In addition to the miRNAs and tRFs highlighted above, several additional species of small RNAs were reported by TEsmall as differentially expressed in the 451Lu BR cells including: siRNAs mapping to transposable element loci, exonic loci, and a variety of structural RNA classes. The structural RNA group included snoRNAs, snRNAs, tRNA fragments, and a vault RNA. The full list of differentially expressed small RNAs can be found in Supplemental File 2.

Comparison of TEsmall with sRNA Analysis Software

Several software packages exist to characterize small RNA data for expression profiling analysis. However, programs designed for this purpose such as miRDeep2 (Friedländer et al. 2012), ShortStack (Axtell 2013), Chimira (Vitsios and Enright 2015), sRNAtoolbox (Rueda et al. 2015), and Oasis 2 (Rahman et al. 2018) typically focus on a particular category of sRNA, predominantly miRNAs. Several packages also consider multiple sRNA types including piPipes (Han et al. 2015), omiRas (Müller et al. 2013), and unitas (Gebert et al. 2017) which include analysis of other noncoding RNAs. However, the output formats of these packages do not lend themselves to easy application of statistical analysis tools like DEseq2 for downstream use and manipulation. While piPipes functions well to annotate and characterize piRNAs by read pileups associated with the ping-pong cycle of piRNAs, it is not particularly suited for annotation of sRNAs from other types of genomic loci, such as miRNAs, siRNAs, and tRFs. PiPipes provides plots of read distribution across lists of transposable elements and piRNA clusters, however, one cannot access tables of these counts with associated TE annotation, suitable for differential expression analysis. While piRNAs are annotated with their respective piRNA clusters, siRNAs are assigned a chromosomal coordinate providing some difficulty in determining patterns in the sources or targets of these reads. It is also of import that intron derived miRNAs like the VIM miRtron were not captured, as there is no mechanism by which to assign siRNA reads beyond mapping the chromosomal coordinates associated to preloaded annotation sets associated with TEs and piRNA clusters. TEsmall, which does not perform piRNA-specific ping-pong analysis, provides a complementary package that is designed to be a general purpose small RNA analysis suite to identify and analyze many types of sRNAs concurrently, presenting the output in a format intended for expression profiling analysis. As piPipes output was not directly comparable to TEsmall output, we performed quantitative comparisons between TEsmall and miRDeep2 output, as seen in Supplemental Figure 2. TEsmall and miRDeep2 preformed comparably with differences originating from higher stringency in TEsmall annotation. This stringency caused fewer reads to be mapped to miRNAs by the TEsmall pipeline. Some reads were attributed by TEsmall to rRNAs and discarded, while others were not attributed to the respective miRNA loci due to mismatched nucleotides within the miRNA. Users interested in the possibility of A-I editing, or other sources of mismatched alignments, may optionally choose to allow mismatches during TEsmall alignment to capture these reads. Output of TEsmall and miRDeep2 annotation was found to be highly comparable with Pearson correlation coefficients of DEseq2 normalized miRNA counts between parental and resistant libraries of 0.882 and 0.910 respectively. Following differential expression analysis by DEseq2 of the TEsmall and miRDeep2 outputs, the Pearson correlation coefficient of log2 fold change values was 0.867. Finally, the novel VIM-encoded miRNA was not captured in the miRDeep2 output. This analysis supports TEsmall as comparable to class-specific sRNA expression analysis packages, while providing information on a wider source of sRNAs.

Discussion

TEsmall is a software package with novel functionality in that it allows the user to simultaneously map and annotate many types of sRNAs including structural RNAs, miRNAs, siRNAs, and piRNAs. This allows one to compare trends in expression between all sRNA types and investigate the cross-talk between distinct sRNA regulatory pathways. Other packages released to date focus on individual sRNA types like miRNAs (Friedländer et al. 2012; Axtell 2013) or piRNAs (Han et al. 2015) and while optimized for these applications, are not adapted for comparison across sRNA categories. In addition to handling multiple classes of sRNAs, the output of TEsmall is formatted for direct integration into downstream analysis pipelines. TEsmall's output files are compatible with statistical analysis software like DESeq2 and efficient heatmap generation. In addition to requiring little data preprocessing, TEsmall outputs an aesthetic HTML file of charts (Fig. 5B) which allows for fast and effortless assessment of library quality, sRNA composition, and size distribution. TEsmall can also be expanded to function for any novel sRNA species provided the appropriate annotation files are available, allowing it to serve as a powerful tool to study RNA biology in many organisms.

We applied TEsmall to a novel dataset in which we compared the effects of BRAF inhibitor resistance on sRNA abundance in melanoma derived cell lines. In this analysis we found several microRNAs whose expression was altered in BRAF inhibitor resistant cells in comparison to parental lines. A table of these hits can be found in Supplemental File 2. Among these candidates, we experimentally validated changes in expression of miRNAs miR-184, -211, and -100. Of particular interest is the novel Vimentin derived miRtron candidate, whose expression pattern was also experimentally validated. Close examination of the characteristic read pile up associated with the VIM miRtron, and secondary structure of intron 6 are all consistent with miRtron processing pathways. Further investigation will be required to determine if this is a true miRtron formed through an intermediate spliceosome derived lariat independent of the Drosha microprocessor subunit, or is instead a canonical Drosha-dependent miRNA.

In addition to revealing miRNAs previously described in the literature, TEsmall detected several novel classes of small RNAs which would not have been found using packages designed for miRNA analysis. TEsmall allows the user to investigate tRNA derived fragments which have been shown to play a critical role in LTR retro-transposon suppression(Schorn et al. 2017). In the melanoma dataset, we identified a novel candidate tRF that appears to derive from ARG-tRNAs and to potentially regulate several HERV-R type LTR elements through occupancy of the primer binding site. Other types of siRNAs that regulate transposon expression were also shown to be differentially expressed in these datasets, suggesting the possibility that transposon-derived transcripts are altered in these BRAF inhibitor resistant melanoma cells.

It is well known that small noncoding RNAs of different subtypes types work in conjunction to regulate cellular processes through complex networks, particularly in the realm of transposon silencing. piRNAs known to regulate transposon expression in the germline have been found to work in cooperation with siRNAs to perform this task.(Tam et al. 2008) In plants, miRNAs have been shown to play a role in transposon silencing by serving as an intermediate to form 21 nucleotide siRNAs via RNA dependent RNA polymerase and while the mechanism would be disparate from plants, hints of miRNAs facilitating transposon silencing have been seen in animals as endogenous and introduced retroviral elements with homologous regions to miRNAs have lower genomic activity(Zlotorynski 2014; Hakim et al. 2008). Current sRNA analysis packages are specific to one or two types of sRNAs making it easy to overlook biologically interesting patterns of interaction between sRNA classes. For this reason, we have created TEsmall, an easy to use package with aesthetic output designed for the concurrent expression analysis of multiple sRNA subtypes.

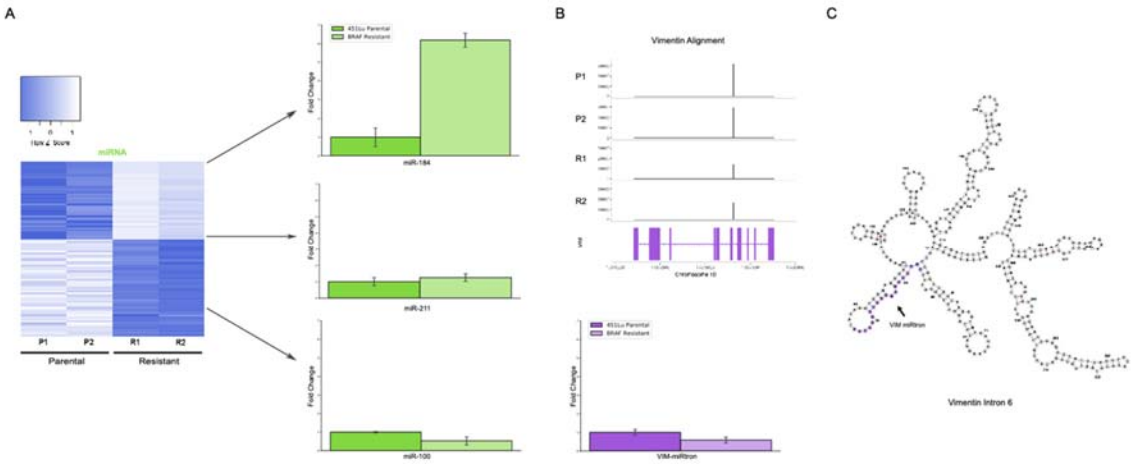


Figure 8 Detailed analysis miRNAs of interest

(A) qPCR representing fold change of miRNAs miR-184, miR-211, and miR-100 in 451Lu BRAF Resistant samples relative to 451Lu parental expression levels across replicates. (B) qPCR representing log fold change of the VIM miRNA in 451Lu BRAF Resistant vs. Parental samples and BAM gene alignment tracks across samples C) RNAfold predicted secondary structure of VIM intron 6 with the candidate miRNA highlighted in purple.

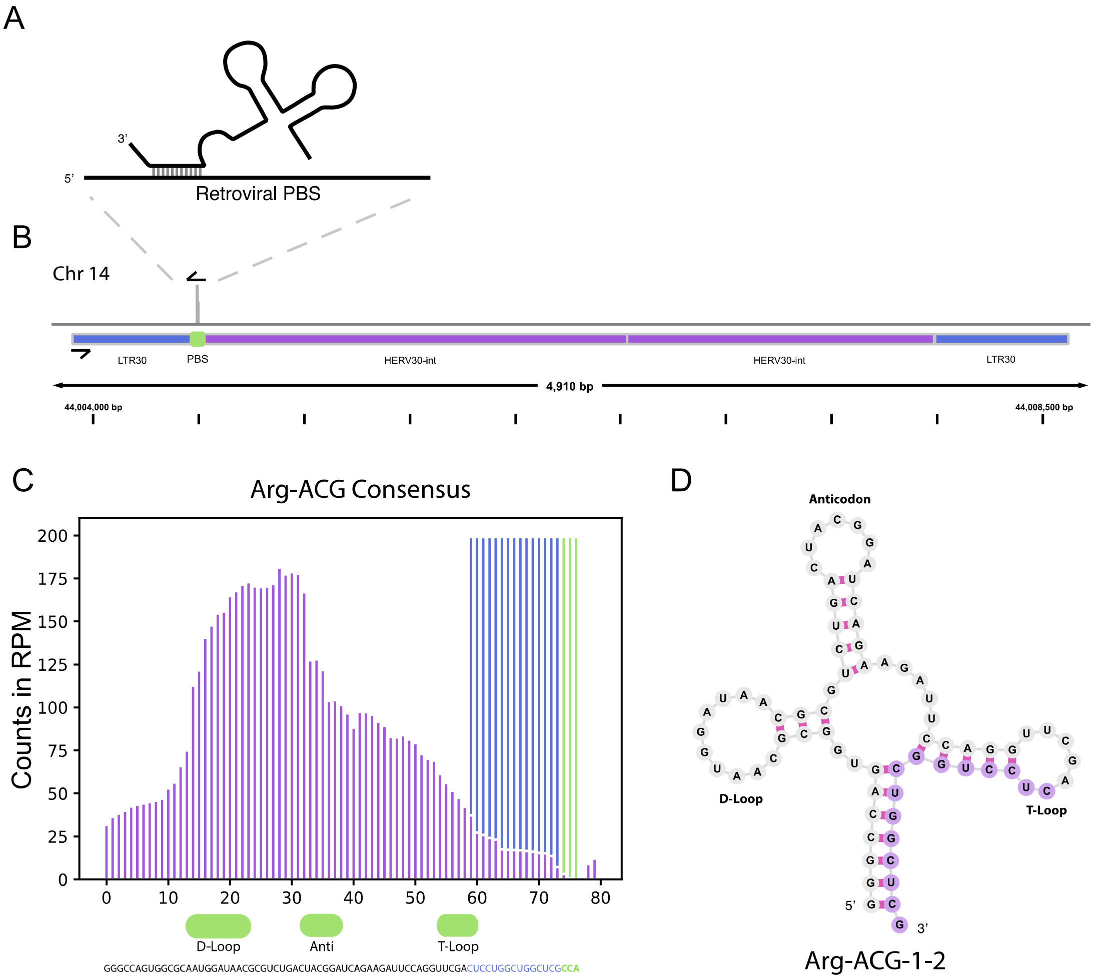


Figure 9 tRNA and TE interaction through primer binding sites

(A) Diagram of primer binding by tRNA to facilitate retroviral reverse transcription. (B) Read alignment track of Arg-CGY family derived 18 nt CCA tailed fragment to HERV30 PBS. (C) Consensus histogram of reads distributed from Arg-CGY tRNAs, and derived 3' tRFs D) Secondary structure of Arg-CGY family member Arg-ACG-1-2 with highlighted 15 nt CCA (-) fragment.

Acknowledgements

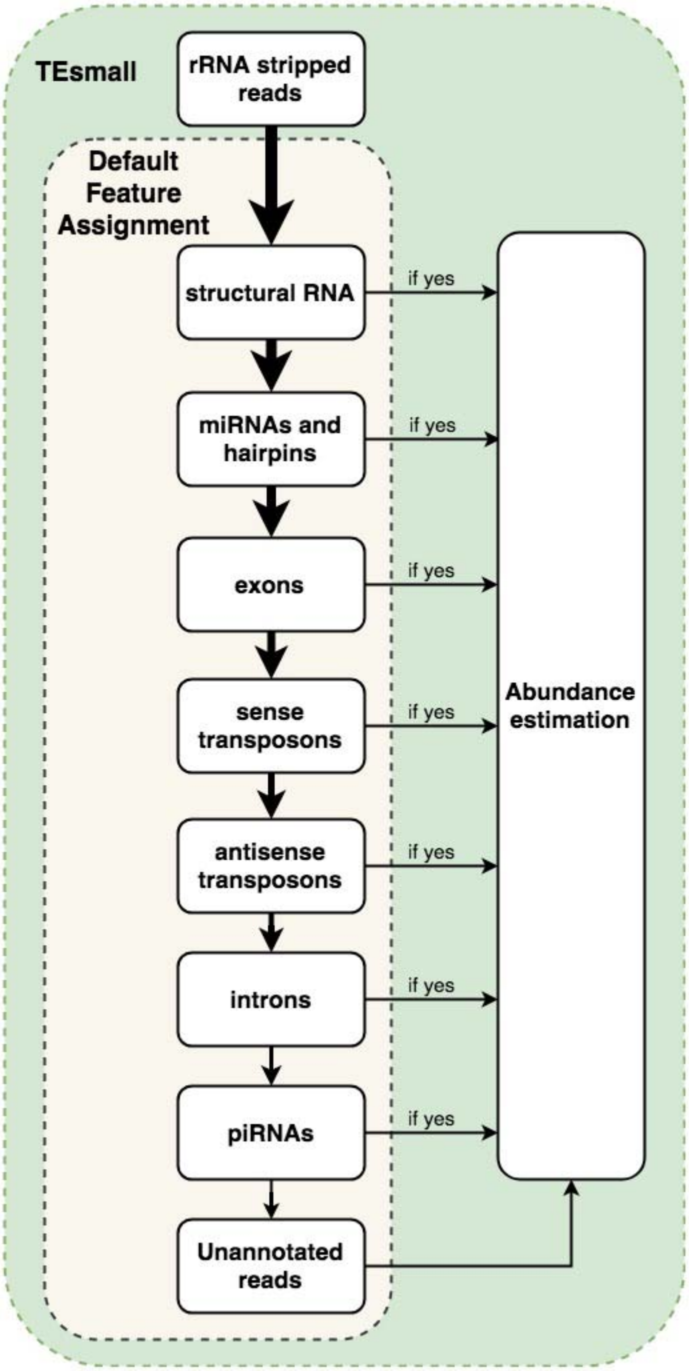
MGH is a scholar of the Rita Allen Foundation. KO acknowledges funding from the NIH training grant 2T32GM065094-16. MGH and WWL were supported by a grant from the NIH/NINDS: 5R21NS088449. We would also like to acknowledge the Meenhard Herlyn lab for help, advice, and for the kind gift of the 451Lu cells.

Author Contributions Statement

Authors KO and MGH analyzed the data, generated the figures and wrote the manuscript. MGH and AP designed the experiments and generated the data. WWL wrote the code for the TEsmall pipeline, packaged the code for GitHub, and contributed to the writing of the manuscript.

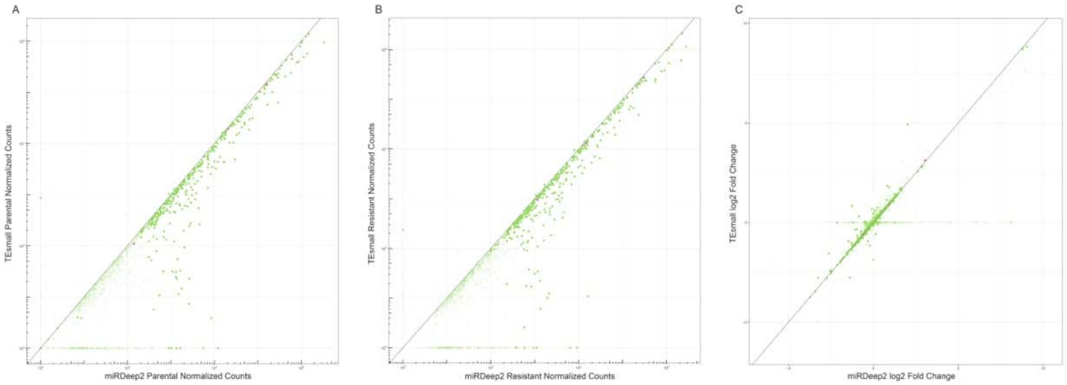
3.2.2 Supplemental figures and files

Additional supplemental files are available at https://doi.org/10.3389/fgene.2018.00461.



Supplemental Figure 1 Flow chart of the default sequential decision tree of TEsmall

Flow chart of the default sequential decision tree used by TEsmall to assign annotations. Alignments are assigned to each category in the indicated order and, if annotated, are removed from the pool before preceding to the next annotation category. Users may opt to re-order the priority table.



Supplemental Figure 2 Scatterplots comparing TEsmall and miRDeep2 miRNA abundance quantification

Mean abundances between biological replicates (panels A-B) and fold change between conditions (panel C) were calculated with DEseq2 on the count tables output by each software package. Low abundance miRNAs with fewer than 2000 counts across all samples are marked as transparent. Shown in pink are the miRNAs validated by qPCR in Fig 4. (A) Log scaled comparison of 451Lu-Par normalized miRNA counts of TEsmall versus miRDeep2, with a correlation coefficient of r=0.882. (B) Log scaled comparison of 451Lu-BR miRNA counts of TEsmall versus miRDeep2, with a correlation coefficient of r=0.910. (C) Comparison of log2 fold change as reported by TEsmall and miRDeep2, with a correlation coefficient of r=0.867.

3.3 Improvements to the TEsmall Pipeline

As described earlier in the chapter, upon publication TEsmall was equipped to handle structural RNA derived sRNAs and siRNAs. However, this pipeline did not take full advantage of the algorithms developed by our lab to handle multimapper ambiguity in transposon derived reads. That is, it did not use EM for statistical redistribution of transposon associated reads, as described for TEtranscripts in Chapter 2. It is known, and reported in the original TEsmall paper, that a significant proportion of the reads in a small RNA library are multimapping, see Figure 10. Therefore, an updated method which integrates statistical methodology designed to handle the multimapper problem was warranted. The original 1/N equal fractional distribution of a read across all mapping locations will reduce the power to call differentially expressed sRNAs and flatten the signal from the true multimapper read stack associated with a small RNA locus. Previously, our lab designed TEtranscripts to better capture the true distribution of multimapper reads from full length mRNA transcripts in a standard bulk RNA-seq library. Yet, it was an open question if the expectation maximization algorithm from TEtranscripts would be able to effectively handle the different distribution of reads associated with a small RNA library. In small RNA libraries, the majority of reads are very tightly distributed in stacks of a single processed sRNA species between 16-36 nt in length, whereas in an RNA-seq library there are pileups of reads across the entire length of a gene or transposon. Therefore, in an RNA-seq library, more information is provided from surrounding distinct multimapper reads for the EM algorithm to use in its probabilistic redistribution process, and it was uncertain if the algorithm would function on such tightly distributed sRNA reads.

I adapted the algorithm from TEtranscripts into TEsmall in a locus specific manner rather than grouped at the level of subfamily, as it is of interest to the field to understand differential expression of species at a single location. A dictionary of unique and multimapping reads were created after the annotation step in the original pipeline for the miRNA and siRNA categories of sRNA transcripts to prevent the redistribution of reads between the sRNA categories. However, the redistribution of miRNA multimapper reads with this algorithm did not perform well and ultimately the scope of the algorithm was limited to siRNAs. In order to test the accuracy of the algorithm, I constructed simulated sRNA libraries with FluxSimulator (Griebel et al. 2012). Full length reads were simulated from a random subset of all TE annotations in the hg19 version of the human genome without any polyadenylation or PCR error. These RNA reads were then fragmented via nebulization and size filtered for fragments within a normally distributed size range, N(22, 3) to represent a population of siRNAs. As these siRNAs were annotated with their exact locus of origin, this was used to determine the ground truth to which all analysis was compared, see Figure 11. In this figure, when comparing the performance of the EM algorithm treated reads to the 1/N treated reads and to the ground truth of the simulation, the EM treated reads are more tightly correlated to the y=x line or a correct read assignment. This shows that the implementation of the EM algorithm did improve performance of TEsmall on siRNA simulated reads. The EM algorithm resulted in a reduction of 34.25% error from simulated siRNA ground truth expression using the 1/N algorithm to 16.5% error using the EM algorithm, cutting the error approximately in half. Percent error was defined as the sum of absolute deviation in counts for each TE instance from ground truth, divided by the total abundance of ground truth counts multiplied by 100. Pearson R^2 values of the 1/N treatment for sense and antisense siRNA reads with ground truth were .976 and .978 respectively and significantly improved to .992 and .993 upon EM treatment. P-values of these comparisons were both approximately 0 using a Fisher-z transformation to test.

In addition to improving the handling of siRNA derived multimapping reads, we were interested in implementing a step in the pipeline which properly handles structural tRNA fragments. The importance of these 3′ tRNA fragments to block primer binding of endogenous retroviruses is described by (Schorn and Martienssen 2018) and in the introduction, and is generally of interest to the field of transposon biology. The original TEsmall paper described the detection of a structural tRNA derived 3′ tRF. However, these required special analysis that was not automated in the original pipeline. While all other tRNA derived fragments are properly handled by the original TEsmall, the 3′ end of tRNA species are post-transcriptionally modified and appended with the three nucleotides "CCA" to serve as the site where an amino acid is attached. As this is a post-transcriptional modification, the presence of these 3′ nucleotides are not genomically encoded, and 3′ tRF reads with the CCA tails are discarded as having too many mismatches to the genome. To address this, unmapped reads are searched for a terminating CCA, the terminating CCA is removed, and the reads are remapped to a tRNA genomic index. Reads which map to a tRNA are labeled and reported in accordance with the rest of the pipeline. This particular functionality of the pipeline does not contribute any particularly advanced statistical methodology, but this type of biologically-aware analysis can have a large impact on the general utility of pipelines, especially for users without strong familiarity with sequencing pipeline construction and adaptation. Figure 12 describes the final construction of the TEsmall pipeline incorporating these changes.

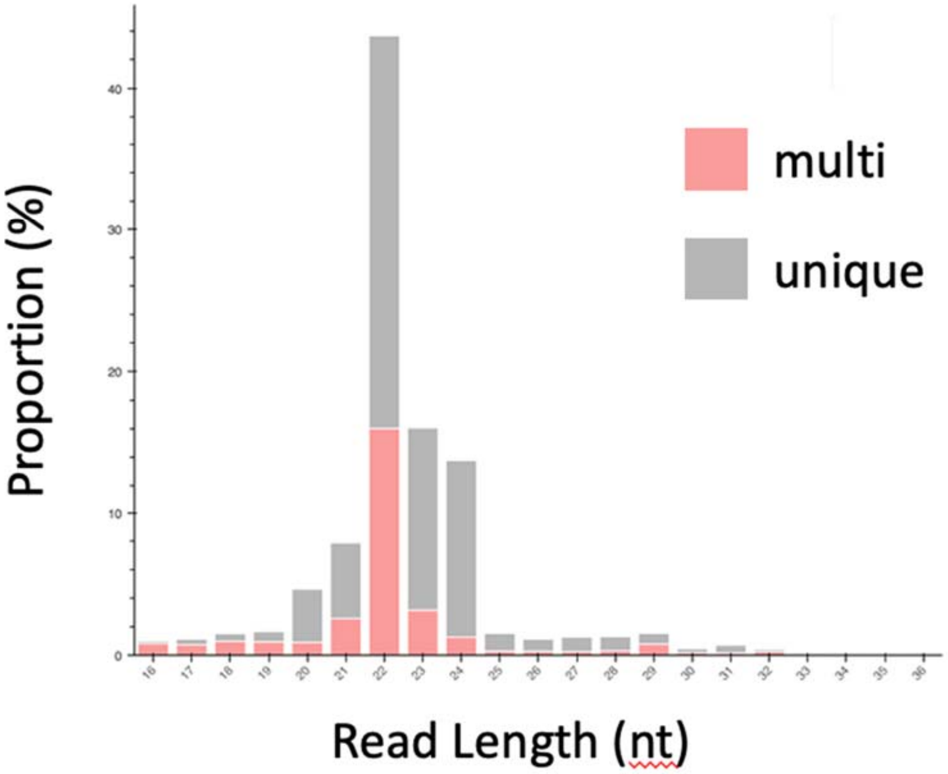


Figure 10 Proportion of multimapper sRNA reads of a certain size in an example small RNA library.

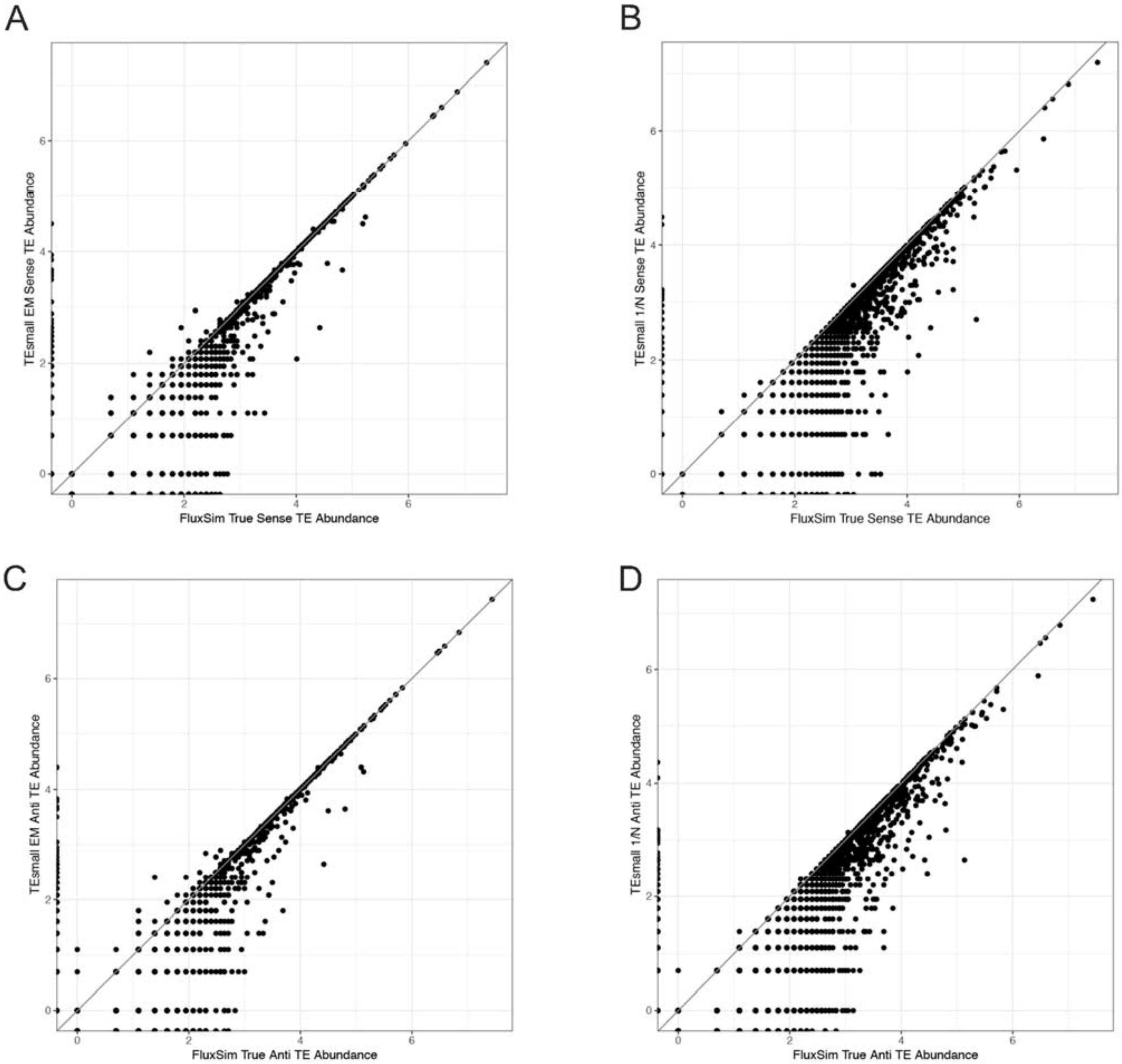


Figure 11 FluxSimulator EM algorithm evaluation.

Plots representing log10 abundance of TEsmall estimated expression levels versus true expression of a FluxSimulator derived TE specific sRNA library. Panels A and C represent EM treated reads with R^2 values of .992 and .993 respectively. Panels B and D represent 1/N normalization with R^2 values of .976 and .978 respectively. This improvement is considered significantly different in the antisense and sense case by Fisher’s-z with a p-value of ~0.

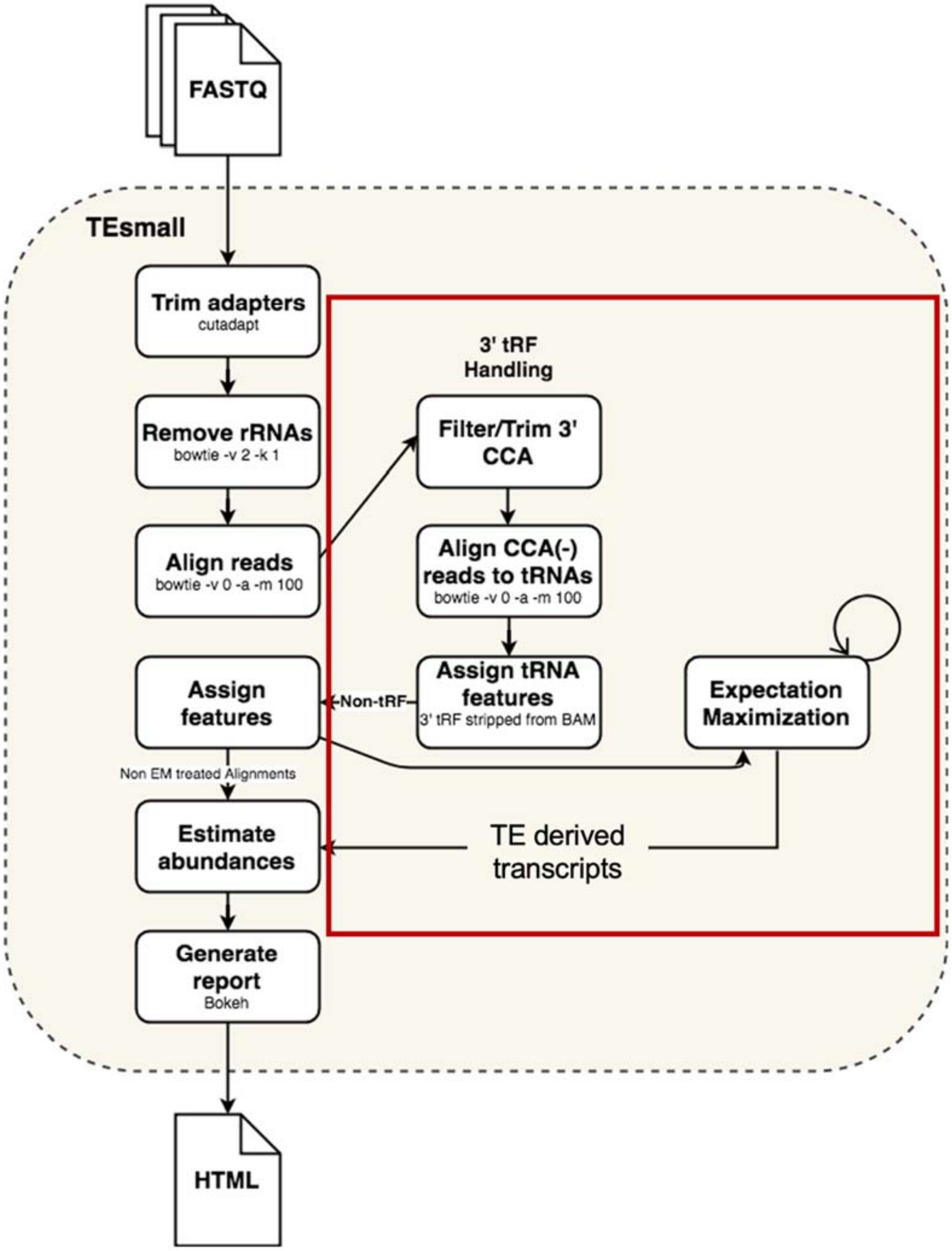


Figure 12 Schematic figure of the improvements to the TEsmall pipeline.

Improvements from the published version are highlighted by the red bounding box.

3.4 Human Adult Somatic TDP-43 loss of function RNA-seq and small RNA-seq analysis

As described in the introduction, TDP-43 aggregation is the primary proteinopathy in amyotrophic lateral sclerosis (ALS). It is still unclear in the field the extent of the effects these aggregates have on the etiology of ALS, and to what degree this aberrant proteinopathy functions as a genetic loss of function, gain of function, or both. I decided to limit the focus of my analysis to the scope of a loss of function perturbation, as it is known that TDP-43 aggregate pathology results in clearing of TDP-43 protein from its primary localization in the nucleus (Suk and Rousseaux 2020). Additionally, our lab had previously observed global TE upregulation in the context of a TDP-43 loss of function perturbation (Krug et al. 2017). This led us to hypothesize TE control might be compromised by inhibition of siRNA production, as siRNAs are well characterized as important for TE suppression. TDP-43 is also known to interact with Dicer, a protein involved in the preprocessing of siRNA transcripts, and supports miRNA processing for those miRNAs which are bound in complex with Dicer and TDP-43 (Kawahara and Mieda-sato 2012). To achieve this aim, a small hairpin knockdown experiment was performed by Nikolay Rozhkov, a former postdoctoral fellow in our lab, to target and reduce the TDP-43 mRNA transcript through RNA interference. This perturbation ultimately reduced the total levels of TDP-43 protein in K562 human lymphoblast cell lines. Lymphoblasts were deemed appropriate for this study as a first investigation as TDP-43 is ubiquitously expressed throughout the human body, and this cell line can survive with a moderate to severe loss of TDP-43 function. The experiment was performed by viral delivery of a short hairpin RNA (shRNA) vector with a guide targeting TDP-43 or a scrambled control. Eight small RNA libraries were prepared in total, 4 respectively for treatment and control, and sent for Illumina sequencing. My analysis began at this point, where I applied the TEsmall pipeline to each of the small RNA libraries, to check the quality of each of the libraries and produce sRNA gene expression tables. Each -library was of high quality and depth, with ~80 million reads and 80% constitution of a 22 nt species representing the miRNAs population. The depth and quality of the sRNA-seq libraries combined with an experimental design with 4 biological replicates leads me to have high confidence in the result of this assay. Additionally, 2 RNA-seq libraries of the mRNA population in the K562 short hairpin knockdown experiment were collected as well with 2 control libraries to analyze the effect changes in the small RNA population might have on potential mRNA targets in the population.

Figure 13 depicts an additional quality control check to the general library analysis described above using principle component analysis (PCA). This analysis reports how similar the variance is between the counts of the libraries prepared. The counts were subjected to variance stabilizing transformation (VST) normalization, a standard count normalization method offered by the DESeq2 package to adjust for library size and log transform to reduce variance in the data (Love et al. 2014). VST normalization is a robust normalization strategy and the primary count normalization method used throughout the remainder of this text for visualization. Upon first analysis with PCA, the largest, or first, principal component separating the sRNA libraries was the strength of the shRNA knockdown. I inferred that this was the major source of variation by noting that the abundance of processed hairpins from our shRNA vectors is captured during small RNA library preparation. This batch effect was corrected by the R statistical package limma (Ritchie et al. 2015) using the removeBatchEffect() command. The successful result of this transformation is depicted in the first panel of Figure 13, as the main principal component separates controls from the treated cells. The original result of the RNA-seq counts did not show a batch effect and did not require correction.

After the libraries were evaluated and processed to ensure proper comparisons were made in subsequent analyses, the resulting counts were compared to look for differentially expressed transcripts between treatment and control in both sRNA and RNA-seq libraries. Differential expression was performed with the DEseq2 package, and the hairpin strength batch effect was included as a covariate in DEseq2's expression model for the small RNA libraries so as not to confound the result of the analysis. As we had seen previously in our lab, transposon derived transcripts were globally upregulated upon TDP-43 knockdown in the RNA-seq libraries as depicted in Figure 14 in the left-hand panel. This is in contrast to genic transcripts, which showed similar levels of up- and down-regulated genes in the same libraries and most transcripts unaffected by knockdown.

An overview of the species which were differentially expressed is depicted in the table in Figure 15. We do not suspect sRNA processing as a whole is affected as there were no global effects on miRNA species and similar numbers of miRNAs were differentially up and down regulated. If global sRNA processing was down as a product of TDP-43 LOF we would expect sRNA production, particularly miRNA production to be down globally. This was not observed, nor did we see any gross changes to sRNA composition across groups between the libraries. miRNAs of interest which were significantly differentially upregulated included miR-18a/b and miR-9-5p. MiR-18a/b was of particular interest as it is known to target transcripts from *DNMT1*, responsible for the maintenance of DNA methylation and *DICER1* a critical protein in sRNA processing. Additionally, *DNMT1* transcript was significantly downregulated in the TDP-43 KD RNA-seq data, indirectly providing supportive evidence that an upregulation of miRs 18-a/b may be post-transcriptionally modifying transcripts important for retrotransposon regulation. *DICER1* was also downregulated, though not significantly. Target transcripts of miR-9-5p, DNMT3A and TET1 were also downregulated in the TDP-43 KD RNA-seq data, with DNMT3A showing significant downregulation. Although, the mechanism of action by which post-transcriptional silencing of these two genes would affect RTE expression in adult somatic cells is unclear as they are generally associated with the de-novo DNA methylation in the early embryo. Regardless, the interplay of these miRNAs and their important targets is an enticing lead for indirect effects of changes to miRNA biology on RTE de-repression. Figure 16 depicts a broader view of the effects of miRNA dysregulation on the pool of mRNA transcripts upon TDP-43 KD. In this figure, a list of high confidence targets of miRNAs were sourced from miRTarBase for all miRNAs which were significantly differentially expressed in the TDP-43 KD experiments. Then, a cumulative distribution function (CDF) was plotted of log fold changes of all miRNA targets between control and KD libraries in the RNA-seq libraries. The mRNAs which were targeted by the significantly differentially expressed miRNAs in the sRNA libraries were plotted in color, and the shift of their CDF from that of the entire mRNA pool was evaluated for statistical significance, using a Kolmogorov-Smirnov test. This was performed for upregulated and downregulated miRNAs separately, and represent the up and right panel of Figure 16 respectively. While, the shift in downregulated miRNA targets was not significant, the shift in the upregulated miRNA target cohort was, and in the appropriate direction. This suggests that the target transcripts of the upregulated miRNAs show anti-correlated expression patterns, consistent with an regulation of those target mRNAs via the TDP-43 dependent miRNAs and providing additional evidence that there may be indirect effects in adult human somatic cells of TDP-43 KD through post-transcriptional mechanisms.

Figure 17 focuses specifically on an interesting effect observed in the TE derived siRNA pool of sRNAs. While there were few individually differentially expressed siRNAs between KD and control, as shown in Figure 15, it was known that RTE transcripts from which these siRNAs derive were in higher proportions in the TDP-43 KD libraries. Normally, the levels of TE-derived siRNAs are strongly proportional to the levels of TE mRNAs, because TE-siRNAs are generated as cleavage products from TE mRNA substrates (Malone and Hannon 2009; Claycomb 2014). However, I observed a change between the proportion of siRNAs being created from their full-length mRNA transcripts in TDP-43 KD cells. Figure 17 is a plot of the shift in these proportion between control and TDP-43 knockdown, where it is observable that the proportion of sRNAs being produced from their precursor TE transcripts is globally reduced upon TDP-43 KD. This phenomenon could be explained by a TDP-43 dependent reduction of processivity of TE derived siRNAs from their precursor transcripts. An alternative hypothesis is that the efficiency of the siRNA production process cannot keep up with the increase in RTE expression associated with TDP-43 knockdown. Experiments to explore these two models for blunted siRNA production in the absence of TDP-43 are currently underway in the lab.

Taken together, these observations of the effect of TDP-43 knockdown in human somatic cells leads us to our current working model of the potential mechanisms by which TDP-43 affects RTE expression through small RNA biology. A schematic of this model is depicted in Figure 18. TDP-43 knockdown can affect RTE levels transcriptionally through increased RTE expression and post-transcriptionally through decreased relative siRNA production. Previously, my lab has shown that RTE transcripts are directly bound by TDP-43 (Tam et al. 2019b), and that TDP-43 chromatin immunoprecipitation data enriches for peaks in TEs (Li et al. 2012). This is evidence for direct transcriptional regulation of TEs.

Data presented in this dissertation provides evidence that there are three additional potential mechanisms by which RTE activation can be modulated post-transcriptionally. The first is indirectly through classical miRNA targeting of transcripts in pathways important for regulation of RTEs, like DNA methylation. The second is through a reduction of processivity of TE derived siRNA transcripts associated with TDP-43 through a mechanism similar to that described by Kawahara and Sato (2012 PNAS). The third is through a more complicated mechanism by which the siRNA processing efficiency of the cell is overwhelmed by higher expression of TEs. Fellow graduate student Craig Marshall is following up the transcriptional regulation of TEs by TDP-43, which was outside of the scope of my dissertation. Fellow graduate student Isobel Bolger is the currently leading the experiments to explore sRNA mediated post-transcriptional regulation of TEs. Isobel and Craig are currently building TDP-43 knockdown systems through CRISPRi which will abate complications in assessing overload of endogenous sRNA processing associated with inducing an exogenous source of highly expressed small RNA precursor hairpins (shRNAs). Isobel will be following up on miRNAs of interest like miR-18 and miR-9 with dual luciferase assays to measure whether these miRNAs have reduced efficiency in targeting in the context of a TDP-43 LOF. Isobel will also be performing AGO2 RNA immunoprecipitation assays in the context of a TDP-43 CRISPRi knockdown to assess the differences to the active pool of sRNAs in adult human somatic cells.

This concludes the section of my thesis describing TDP-43 loss of function in adult human somatic tissues. In the next chapter I will describe a separate but related topic of classifying and exploring transcriptional subtype in the neurodegenerative disease ALS. Our lab has previously described RTE activation in one of the three subtypes of ALS, and has associated it with the accumulation of phosphorylated TDP-43 in the cytoplasm. It is my hope that future work from our and many other labs will eventually bridge mechanistic insights about the normal function of TDP-43 with insights into the etiology of the disease TDP-43 associated disease, ALS.

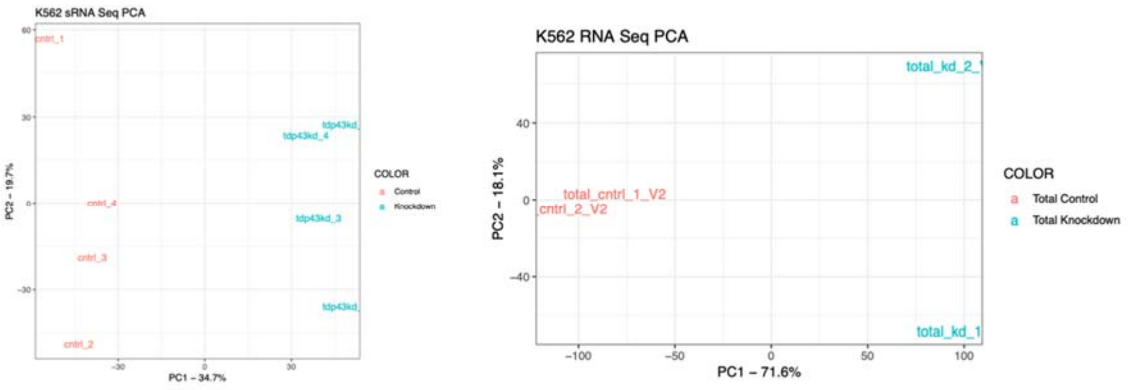


Figure 13 Quality control PCA plot of sRNA and RNA sequencing libraries.

These plots depict variance of normalized counts between multiple sRNA and RNA sequencing libraries after batch correction for strength of knockdown. The percentage of the variance described by a principal component is reported next to the label. The left panel shows in the sRNA libraries, the greatest variance falls across control vs knockdown after batch correction which is ideal. The right panel shows the same characteristic across the standard RNA-seq libraries.

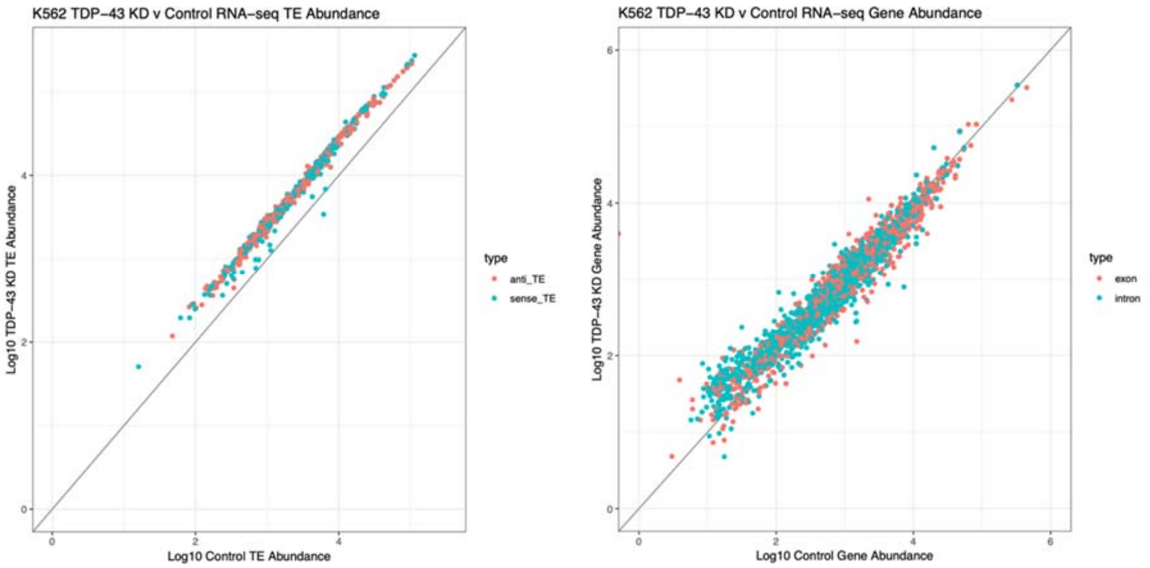
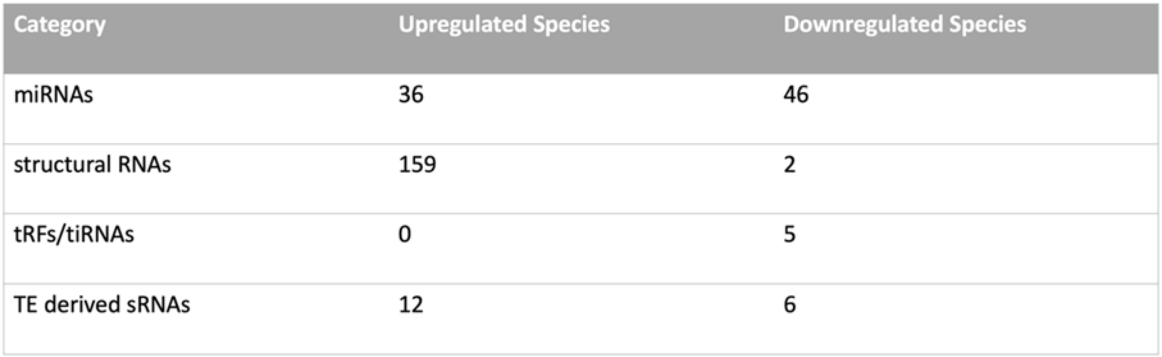


Figure 14 Scatterplots of TE and genic abundance after TDP-43 knockdown.

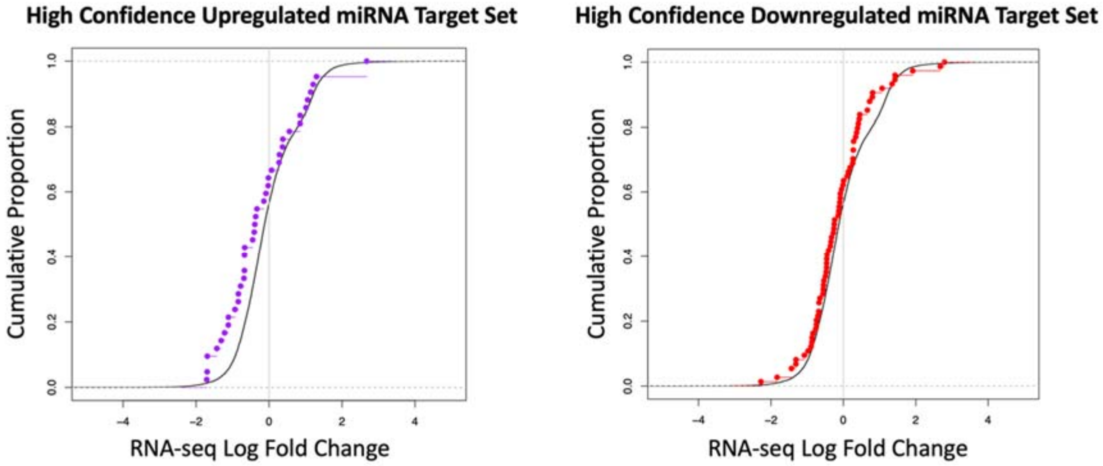
Scatterplots depicting that transposon derived reads are globally upregulated in K562 RNA-seq TDP-43 knockdown libraries (left panel), while genic derived transcripts vary symmetrically (right panel).



Category	Upregulated Species	Downregulated Species
miRNAs	36	46
structural RNAs	159	2
tRFs/tiRNAs	0	5
TE derived sRNAs	12	6

Figure 15 Differentially expressed sRNA species upon TDP-43 knockdown

A table outlining the number of species significantly differentially expressed with respect to control within each small RNA category.



Komolgrov-Smirnov Test: D = 0.232, p-value = 0.02*

D = 0.104, p-value = 0.4025

Figure 16 Cumulative distribution plots of log fold change between TDP-43 knockdown RNA-seq libraries and control

Colored circles mark high confidence miRNA target genes of miRNAs significantly differentially expressed in our small RNA libraries. High confidence targets are defined in miRTarBase as validated by western blot or luciferase assay as direct targets. Left panel: Targets of significantly upregulated miRNAs in K562 small RNA TDP-43 knockdown libraries. A significant shift to the left is detected capturing the hypothesized effect of TDP-43 dependent upregulated miRNAs downregulating their targets. Right panel: Targets of significantly downregulated miRNAs. No significant shift indicating the hypothesized upregulation in targets is detected.

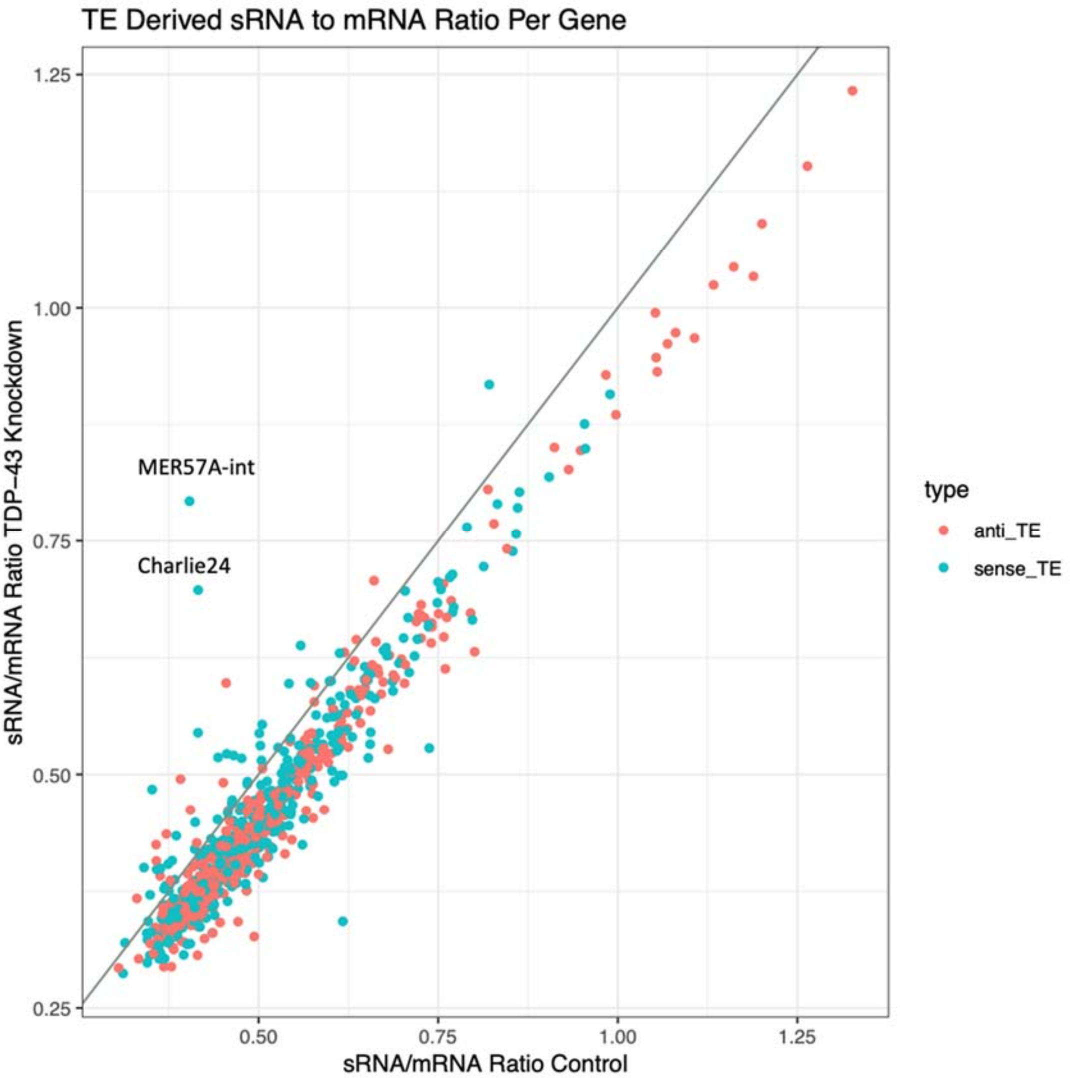


Figure 17 A scatterplot depicting the ratio of siRNAs to their originating TE transcript in TDP-43 knockdown libraries versus control

The ratio of sRNAs to their originating transcript is reduced in TDP-43 knockdown libraries as most data points fall below the line y = x depicted in grey.

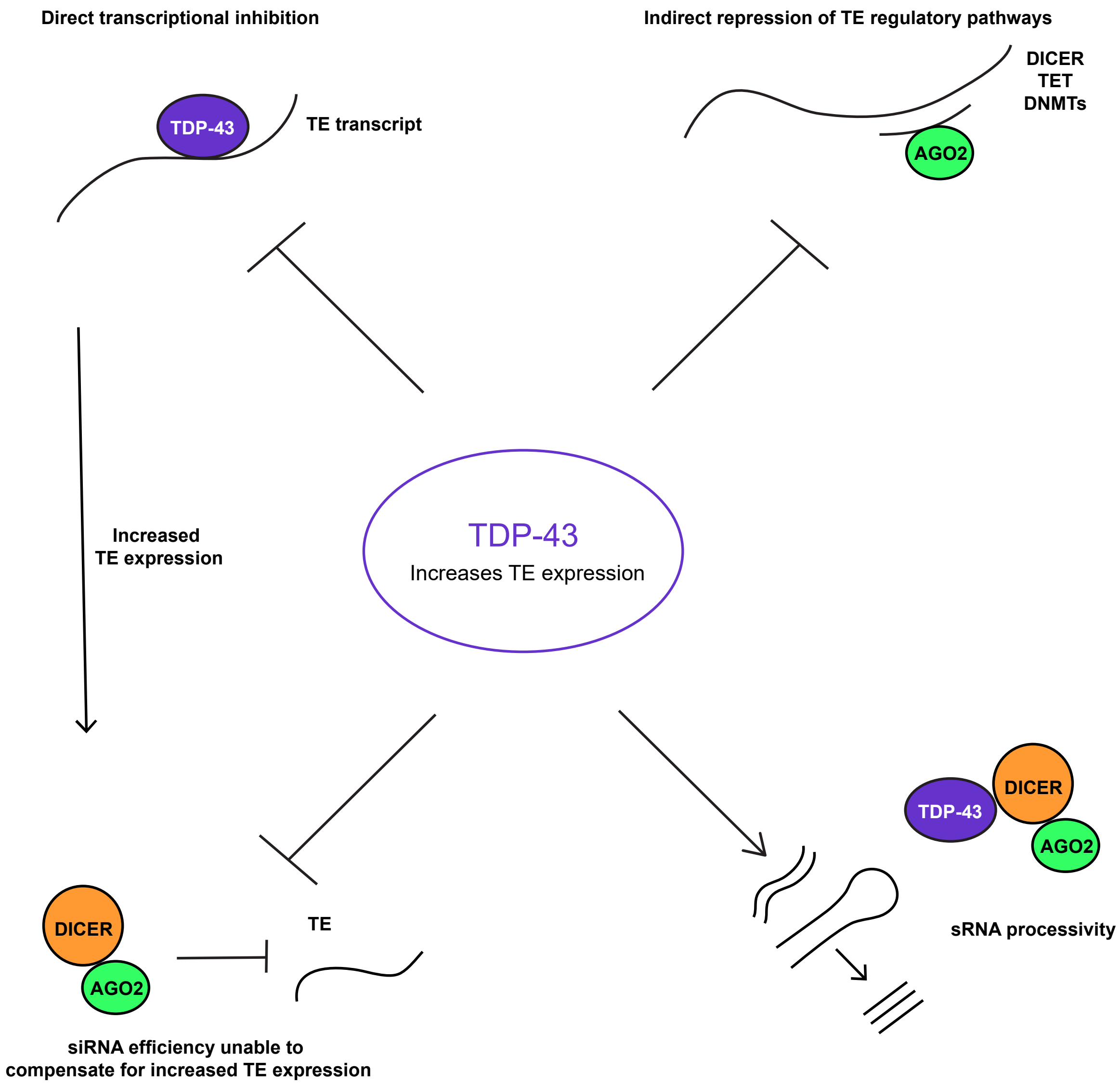

Direct transcriptional inhibition
TDP-43
TE transcript
Indirect repression of TE regulatory pathways
DICER
TET
DNMTs
AGO2
Increased
TE expression
TDP-43
Increases TE expression
TDP-43
DICER
AGO2
sRNA processivity
DICER
AGO2
TE
siRNA efficiency unable to
compensate for increased TE expression

Figure 18 A diagram of the multiple putative models contributing to TE upregulation upon TDP-43 loss of function in adult human somatic cells

In a clockwise direction from the upper left: TDP-43 is a direct transcriptional repressor of TE derived transcripts through its role as a transcription factor (data not shown). TDP-43 indirectly represses TEs through modulation of their regulatory factors like small RNA processing proteins or DNA methylases. TDP-43 increases processivity or stability of sRNA molecules important for TE control. Finally, at the bottom left, TDP-43 loss of function results in a cell state wherein sRNA TE control is overloaded and cannot compensate for the abundance of TE derived transcripts.

Chapter 4: A classifier for ALS/FTD-TDP subtype and investigations

4.1 Rationale for development of a classifier of ALS/FTD-TDP subtype

In the previous chapters I discussed tools to profile transposon biology, and how these tools are important as TE biology is relevant in the study of human disease, particularly ALS. In this chapter I will be focusing on expanding upon our original research into the molecular subtypes of ALS. ALS subtypes were defined as disparate transcriptional states observed in the frontal and motor cortex tissues of ALS and FTD-TDP patients *post mortem* (Tam et al. 2019b). Tam et al. identified these subtypes in a preliminary cohort of 176 samples from 95 patients, 77 ALS patients and 18 controls, obtained as part of the New York Genome Center (NYGC) ALS Consortium. This pilot ALS Consortium cohort study, which was the largest ALS patient genomics profiling study when released, laid the groundwork for the definition of molecular subtype in ALS, and found three predominant groups that dominated the transcriptional profiles in separate subsets of ALS patient tissues. The most frequently identified was the 'ALS-Ox' group, or a group with upregulated pathways associated with oxidative stress. The 'ALS-Glial' group was associated with activated microglia and immune signaling pathways. The 'ALS-TE' group, was associated with upregulated TEs and the presence of dense phospho-TDP-43 cytoplasmic protein aggregation. These subtypes were defined using the non-negative matrix factorization method, or NMF, which is a de-novo clustering algorithm, the mathematics of which was described in the introductory chapter. In this application of NMF by SAKE (Ho et al. 2018), the optimal number of clusters (*k*) can be found by simultaneously maximizing the cophenetic correlation scores and the silhouette score of the clustering across several values for the number of clusters. Silhouette scores measure the mean within-cluster distances as compared to the distances to members in the next-nearest cluster. Cophenetic correlation scores are meant to reflect how well a particular clustering reflects the actual pairwise distances between any two samples and is defined by the relative correlation distance of any two cluster members and the distance that must be traveled up a particular clustering tree in order to join those two points. The NMF algorithms, as implemented by SAKE, is typically iterated 500-1000 times, randomizing the order in which samples are assigned to clusters, in order to ensure that the final clustering result is a consensus mean of many calculations of the optimal clustering of all samples. The NYGC ALS Consortium samples were found to have an optimal clustering of k=3, which separated the ALS-Ox, ALS-Glia, and ALS-TE groups.

Characteristics of the ALS subtypes were defined through differential expression analysis by DEseq2 between clusters and control samples, and pathway enrichment analysis. Additionally, these transcriptomic signatures were validated by staining for protein markers of the identified dysregulated pathways, using immunohistochemistry of matched tissue sections from 40 patients in the NYGC ALS cohort. To validate the ALS-Glia group, patients were stained for IBA1, a marker of microglial activation. To validated the ALS-TE group, and its association with TDP-43 dysfunction, patient tissue slides were stained using a pTDP-43 antibody to quantify TDP-43 aggregation pathology. Using IHC validation, we were able to verify that the ALS-Glial samples did have both transcriptional signatures of glial activation and neuroinflammation as well as elevated protein markers of glial activation, but not a simple loss of other cell types. Similarly, we were able to demonstrate that the ALS-TE subtype samples with elevated TE expression were the most likely to have dense aggregation pathology. This provided the baseline of our understanding of ALS subtypes. However, NMF is a de-novo unsupervised clustering algorithm, not a supervised classification algorithm, and is such so not suitable for classifying new samples as they are added to the ALS Consortium cohort. This chapter describes the creation of a classification algorithm to allow for the facile assessment of incoming patient samples with the potential to expand to additional tissue/data types. Along the way to creating a deep layer neural net for ALS subtype classification, I also provide analysis on the information encoded within the classifier through WGCNA modules and some additional analysis with external data to better understand ALS subtype.

4.2 A gaussian mixture model as a classification algorithm

We know that the landscape of ALS subtype is a transcriptional mixture of disparate molecular pathways. I looked to see how the transcriptional expression was distributed across patients and found that it was largely consistent with a gaussian distribution, with a large population size of n = 176. Therefore, I hypothesized that a gaussian mixture model (GMM) may be an ideal classification algorithm to train and use to predict subtype on incoming samples. The original NMF method performed feature selection using median absolute deviation (MAD). This MAD thresholding selected the top 5000 most variable genes for input into the clustering algorithm, after removing sex associated genes. Sex associated genes were defined by DESeq2 (Love et al. 2014) differential expression analysis for genes that were significantly differentially expressed across sex. I used this same feature selection for input into a GMM through Scikit-learn (Pedregosa et al. 2011) and performed a training regime on the original Tam et al. cohort using an 80/20 train-test split. The test cases were chosen to be the hardest cases, or those with the lowest probability of assignment as described by NMF. The classification labels which the algorithm used to train on were the original NMF labels of assigned ALS subtypes. The model was determined to be highly overfit as it ascribed a probability of assignment of 1, or 100% confidence on label assignments to each sample, but the GMM assigned labels did not agree with the original NMF ALS subtype labels as depicted in Figure 19.

The overspecification of the GMM led me to recognize that direct classification of ALS subtypes in a high dimensional gene expression space was a very difficult problem for a classification algorithm to navigate. In the field of machine learning, there is always a tradeoff between bias and variance in any model. The bias of the model is the assumptions a model is making about the data to make it easier to approximate its structure. The variance of a model is how adaptable (or robust) the model is to variation in the data. In the case of the GMM, the model's variance was too high for the features it was provided over the size of the dataset, leading to overfitting. It was at this point that I began to look into more clever feature selection methods, beyond MAD, as preprocessing for a classification algorithm. The hypothesis was that encoding the gene expression data into a lower dimensional manifold would denoise the transcriptional data. This denoised lower dimensional embedding would then allow the signal of subtype to be characterized more readily. Additionally, I sought a dimensionality reduction method which would integrate well with a classification algorithm downstream. This facility of integration is what led me toward neural networks as their architectures are extremely flexible, and there are many types which perform classification and/or lower dimensional embedding.

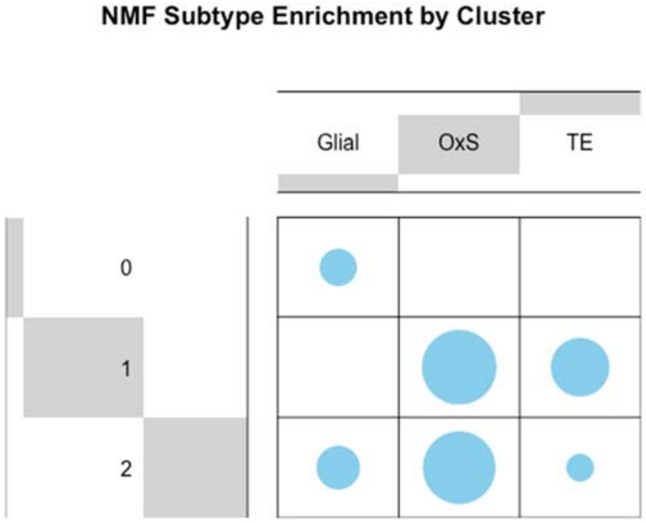


Figure 19 A balloon plot of the confusion matrix between assignments from a trained gaussian mixture model and NMF based cluster assignments

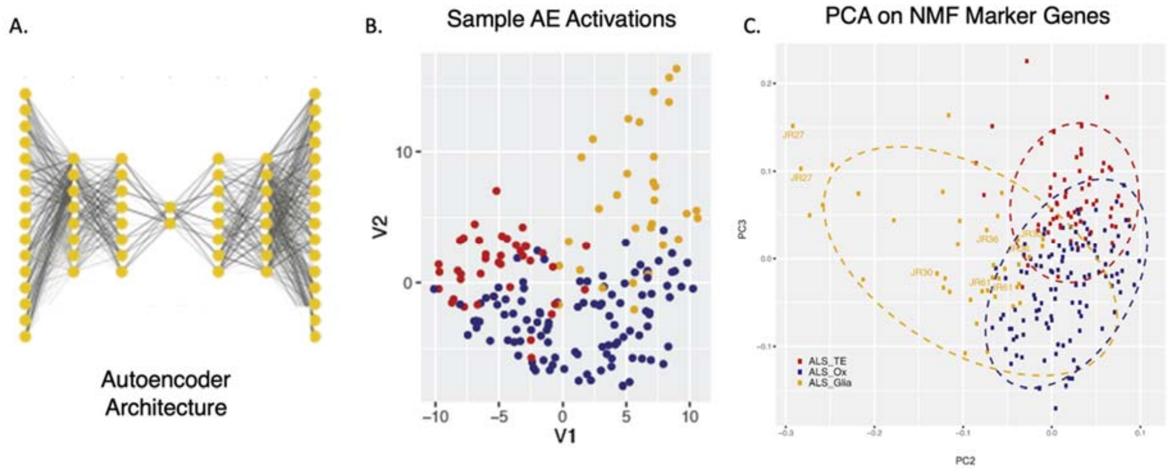
This depicts that the gaussian mixture model was unable to effectively capture NMF subtype as many cases are off of the identity matrix (or off-diagonal).

4.3 Neural networks to facilitate and perform classification

To first address how well a neural network could embed the Tam et al. (Tam et al. 2019b) transcriptional dataset into a lower dimensional manifold, I trained an autoencoder to compress the 5000 MAD gene list from the original study. As described in the introduction, an autoencoder is a neural network which takes data of interest into an input layer the size of your dataset and then bottlenecks, or encodes, that data through layers of decreasing size into an innermost layer of a desired lower dimensionality. This innermost layer is then expanded and decodes that dataset back to its original form. The model is trained by minimizing the mean squared error (MSE) between the input and output layers, thus creating an internal embedding in the smallest layer which can most faithfully reproduce the original dataset during decoding. The neural activations of each sample from this innermost layer can be visualized or used as inputs for downstream analysis, such as classification. I successfully trained an autoencoder on the original NYGC ALS Consortium dataset, with the results displayed in Figure 20. In Figure 20, the two axes represent the activations from the innermost two nodes of the autoencoder – these are the embeddings of the Tam et al. dataset that the autoencoder learned, when trained to find the optimal two-dimensional encoding of the entire list of 5000 MAD genes across 176 samples. The colors in Figure 20 represent the original NMF labels for ALS subtype, which are cleanly separated in this autoencoder defined space. It is important to reiterate that the NMF labels were never included in the training of the autoencoder; it was merely designed to encode a 2-dimensional embedding of the 5000 MAD genes in the previous study. Although the lower dimensional embedding in Figure 20 is not used for classification in this figure, as an autoencoder is not a classifier, the ability for the activation nodes to separate samples by subtype is striking and reassuring evidence of the robustness of the signal for molecular

ALS subtype in the data. It is important to note that ALS subtype was originally identified by a very different computational method (NMF), which was also able to naturally pull apart the signature of ALS subtype as the predominant biological signal in the data. To demonstrate the similarity of the results returned by the autoencoder and NMF, comparison of Figure 20 panels B and C shows that the lower dimensional embedding of the autoencoder (Figure 20-B) shows the same relative distribution of sample similarities as a PCA plot of the samples by NMF marker gene list (Figure 20-C).

Since the neural network architecture was able to separate molecular structure that corresponded to ALS subtype, it occurred to me that I could use this embedding for classification. I designed a multilayer feedforward perceptron (a particular form of neural network classifiers) which would connect the encoder from a trained autoencoder as an input into a neural network designed for classification. The weights of the autoencoder would be frozen in the training of the classifier so that the embedding would not be perturbed, and the neurons in the classifier would be trained to draw a decision boundary for NMF classification on this lower dimensional manifold. Unfortunately, this first approach was not successful in producing a robust classifier, returning a maximum classification accuracy of 75%, as depicted in Figure 21.



Tam et al. 2019

Figure 20 Subtype emerges from an agnostic embedding of the top 5000 median absolute deviation selected genes used in NMF subtype clustering via an autoencoder

Panel A depicts a schematic representation of the autoencoder architecture. Panel B is a scatterplot of the values of the activations of the two innermost nodes of the autoencoder architecture and is colored according to the original NMF labels. Panel C is a scatterplot of two of the top principal components of the NMF marker gene list, which is a list returned by NMF of the genes most strongly associated with each NMF cluster assignment; these points are colored by their NMF subtype assignment and the colors are matched to those in panel B.

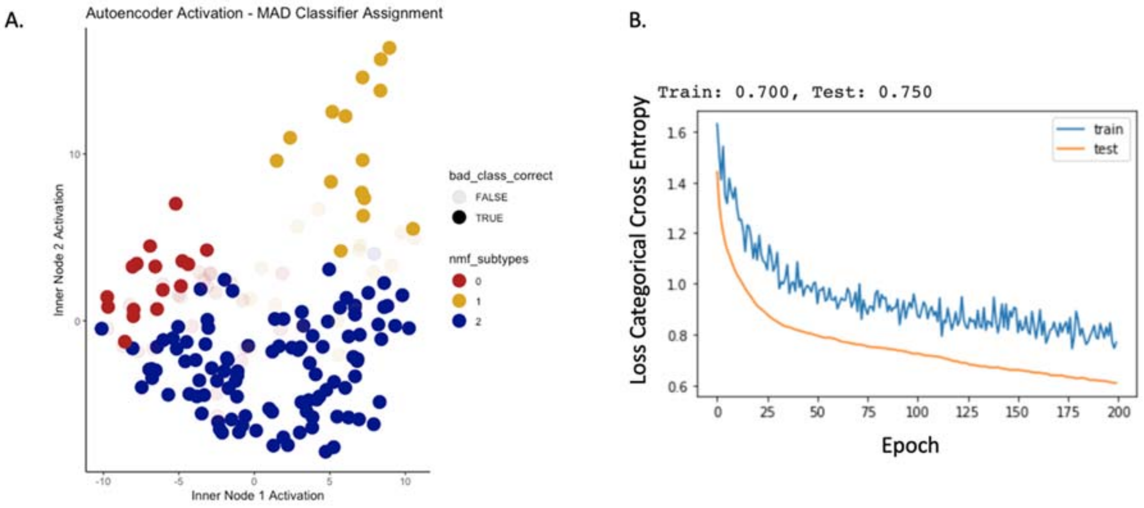


Figure 21 A figure depicting the performance of a MAD 5000 autoencoder based classifier

Panel A depicts the correctly called ALS subtype samples in bold colors according to their subtype lables (ALS-Ox in blue, ALS-Glia in yellow, and ALS-TE in red) on the same autoencoder activation plot featured in Figure 20. Incorrectly called samples are shown as translucent colors and predominantly occur along the boundaries between clusters of ALS subtype samples. Panel B is a plot of loss by categorical cross entropy of the samples in the training and test set after each epoch of training. The final accuracy of predictions on the test and training sets from this classifier are depicted in the title of this panel at 75% and 70% respectively.

For training this initial autoencoder-based classifier, in each training protocol an 80/20 train-test split was used, specifying that 80% of the data was used to train the classifier, while 20% of the samples were reserved for testing the trained classifier's ability to correctly identify the ALS subtype of a particular sample. Several differently sized architectures were trained with the optimal size selected based on performance. All neural networks described in this thesis were built with the keras package (Chollet and Others 2015) which is a wrapper around TensorFlow, version 2 (Abadi et al. 2015). Performance for the autoencoder was initially defined ad hoc for ease of visualization (i.e. a 2 dimensional encoding). As different numbers of inner nodes were attempted to find more optimal node/layer architectures, these were evaluated by the greatest mean squared error loss for additional inner nodes added.

The optimal MAD-based autoencoder trained for classification of ALS subtype had a final architecture of 5000-265-256-6-256-256-5000 nodes – in other words: an input layer of 5000 nodes to represent the 5000 MAD genes, two fully-connected layers of 256 nodes each, and an inner bottleneck layer of 6 nodes, which was mirrored for the output decoding layers. The autoencoder used for visualization had the same architecture with an innermost layer with 2 nodes instead of 6. In addition to the layers of densely connected nodes with ReLu activations, after each densely connected inner layer, I added a BatchNormalization layer, which keeps the mean activation around 0 and fluctuation of activations within 1 standard deviation, and a Dropout layer, which randomly sets inputs to that layer to zero with some frequency as a form of regularization to prevent model overfitting. The dropout rate I chose and used in the entirety of this thesis was 50%. Dropout is not used for prediction; it is only employed during training while weights are being updated. The loss function for autoencoders was mean squared error (MSE), and the training was automated using early stopping. While using early stopping, a performance metric is recorded after each training epoch, a forward and backpropagation step; after a certain number of epochs without improvement of the chosen metric (MSE), training is terminated. This is another method used to prevent overfitting of the model as often a model is fit before the number of prescribed training epochs. In this case, the metric monitored was mean squared error, and patience was assigned to 20, so that the training regime would wait for 20 epochs without decrease of MSE before terminating training. In all training regimes throughout this thesis, the maximum number of training epochs was 200.

Once the MAD-based autoencoder was optimally trained, this was used as input to a feed forward classifier with very similar architecture: 5000-256-256-6-8-8-3. This depicts the front encoder from the trained autoencoder, attached to two densely connected ReLu layers with BatchNormalization and Dropout, and a final softmax output layer which reports a probability of classification to each of the 3 ALS subtypes. Early stopping was also employed with a patience of 20. While the initial results using the autoencoder to visualize the separation of ALS samples by ALS subtype in its innermost bottleneck layer were very promising, as shown above, this particular classifier performed relatively poorly. Moreover, this classifier could not be improved by increasing the number of input nodes in the bottleneck layer, nor by increasing the number of nodes in the densely connected inner layers, nor by increasing the number of densely connected layers. Next, I will describe how preprocessing the input data by weighted gene correlation network analysis (WGCNA) ameliorated this problem.

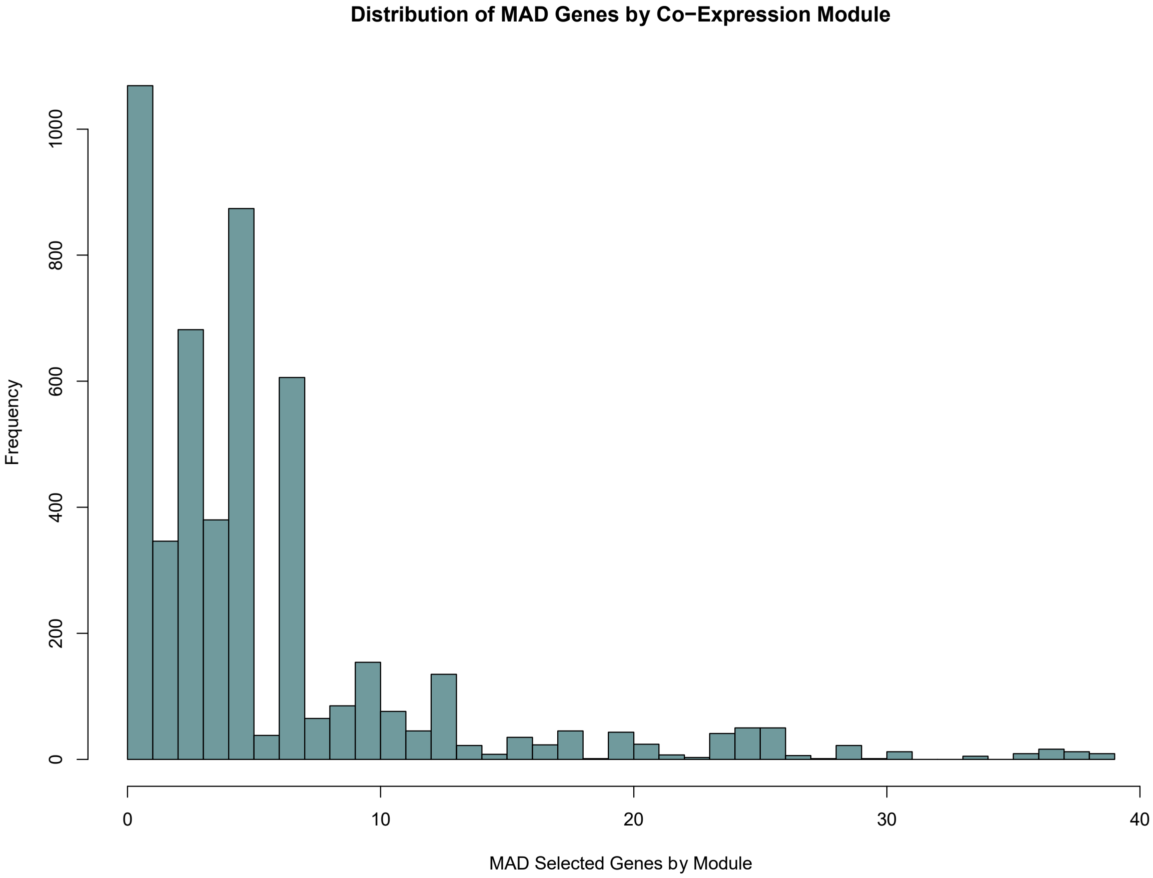


Figure 22 A figure depicting the number of MAD genes within each WGCNA module

This figure visually represents the biased sampling of MAD genes within larger co-expression modules. Some modules of correlated/co-expressed genes were represented hundreds to thousands of times in the MAD list, while other modules had low representation or were missing entirely from the MAD list. Through using WGCNA, this distribution is flattened and the eigengene value from each module is input with equal weighting.

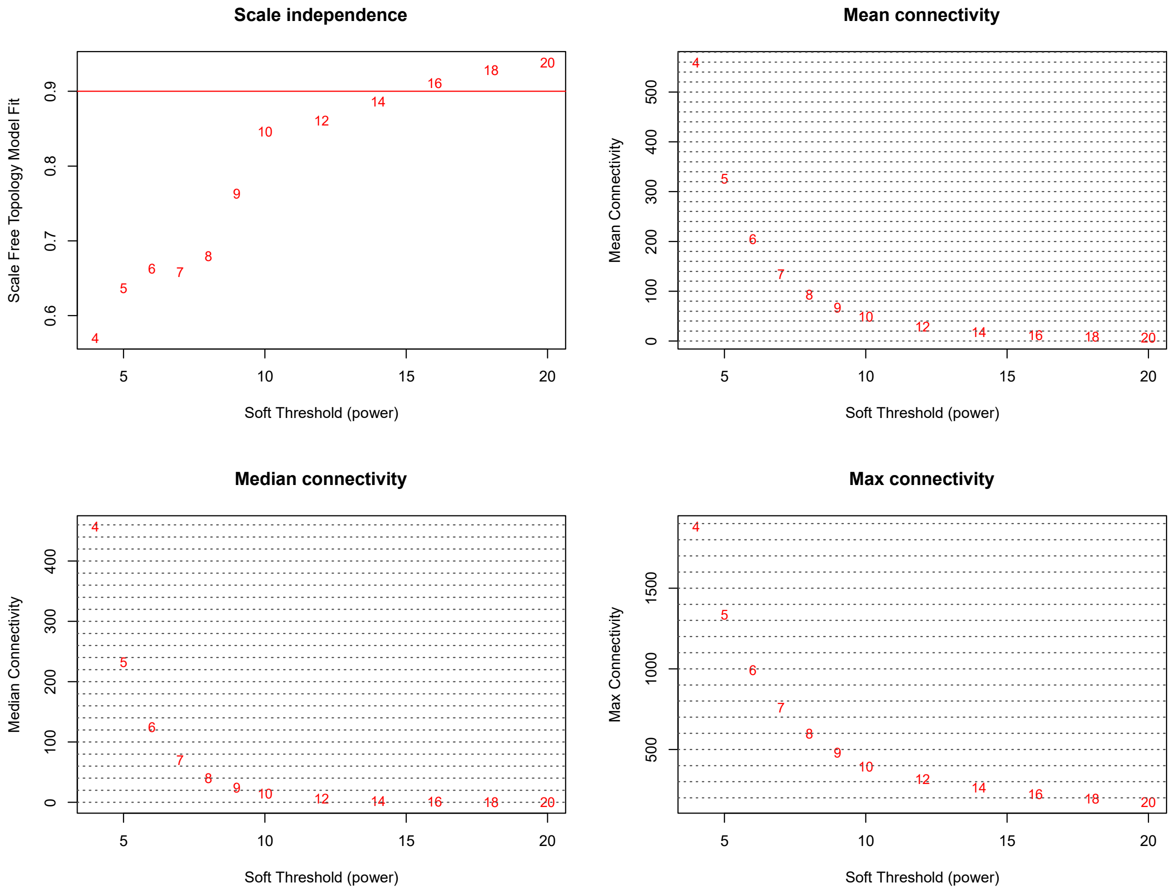


Figure 23 Multiple selection criterion to determine the exponential/power parameter for the decay of the adjacency function in WGCNA

The upper left panel plots the R^2 Pearson correlation between log(p(k)), where p(k) ~ k^{-power} is the connectivity of a particular fit adjacency matrix, and log(k) where k is the connectivity of a scale free network for a set of user defined power values. A value of 1 would indicate that the network perfectly satisfies the scale free criterion, but in practice a R^2 value of ~.8 is satisfactory. For this analysis we chose a power parameter of 9 as it is near 0.8 and maintains a higher connectivity within the network. The remaining panels of the figure show the summary statistics of the connectivity of the network for each user defined power parameter. Connectivity is defined as the sum of the connection strength/adjacency of a gene with all other genes in the network.

4.4 WGCNA as a feature selection method for classification by neural networks

Weighted gene correlation network analysis (WGCNA) (Langfelder and Horvath 2008) is a method in which genes are grouped into co-expression networks by their relative correlation across samples in a given dataset. WGCNA assumes that gene co-expression networks are organized according to a scale-free topology. The scale-free assumption asserts that genes are not randomly associated with each other; rather genes tend to organize in hubs (clusters with similar expression patterns), and the connectivity of these hubs is distributed according to a power law, as described in detail in the introduction. The biological relevance of this assumption is that genes are not randomly expressed, but rather are co-regulated in cohorts by sets of common signals like transcription factors. In short, co-regulated genes will have similar co-expression patterns across samples.

To construct a WGCNA network for a given dataset, the user must select the power parameter which best fits the decay function in the connectivity of the network. This is optimized by looking across several descriptive statistics of the network (e.g., mean, median, and max connectivity) and choosing the threshold which optimizes the network. In the case of the Tam et al. data, the optimal threshold was 9, and the descriptive statistics can be visualized in Figure 23 as calculated and described in (Zhang and Horvath 2005). The network was constructed on the entire gene list associated with the Tam et al. dataset, after filtering for expressed genes defined as having an expression level higher than a mean of 10 counts, leaving ~36,000 genes which were VST normalized by DEseq2 (Love et al. 2014). This resulted in a network with 40 modules, one of which was composed of sex associated genes and was excluded from downstream analysis. WGCNA is being used, in this context, as a dimensionality reduction method such that 36,000 genes are now being represented by 40 WGCNA modules.

In addition to its utility as a dimensionality reduction method, WGCNA also overcomes a problem inherent in using median absolute deviation (MAD) as a feature selection method. More specifically, the sets of highly variable genes above some MAD threshold might not evenly sample across all co-expression networks. Some co-expressed gene modules might be quite large and highly variable in a given dataset, together enriching the MAD list. Other co-expressed gene sets might be smaller and underrepresented in the MAD list, but nevertheless highly informative. This is exemplified in Figure 22 where the distribution of the number of genes in the Tam et al. 5000 MAD list for the NYGC ALS Consortium dataset is shown over each WGCNA co-expression module for that same dataset. WGCNA overcomes this skewed distribution problem by flattening the size of a gene co-expression module into a single value, called an eigengene. An eigengene is defined as the first principal component of the expression of all of the genes in a co-expression module.

The relationship of these WCGNA modules to the variance of the data with respect to ALS subtype is depicted in Figure 24, where the NYGC ALS Consortium samples are plotted for the first two principal components of the WGCNA module eigengenes. Here, WGCNA module eigengene values are interesting in their own right. They allow us to investigate which sets of genes are co-expressed, and how these co-expressed genes are differentially acting with respect to subtype, which will be discussed further near the end of this chapter.

At this point, we will focus on how these eigengenes can be used as inputs to represent the transcriptional landscape in a 39-D space (40 modules minus the sex-linked module). The value for the eigengene of each sample was computed and this value was used as the input for a standard classifier very similar to the one described in the first attempt above. The classifier was trained exactly as described in the previous section with an 80/20 train-test split. The final optimal architecture chosen was a neural network with 39 input nodes (one for each WGCNA module eigengene), 6 fully connected inner nodes, and 3 softmax outputs for each of the 3 respective ALS subtypes. This network was not trained using Dropout or BatchNormalization layers as it was not a particularly deep network. A schematic of the network along with the training profile is featured in Figure 25. Surprisingly, this very simple neural network architecture, combined with the WGCNA preprocessing of input data, increased the classification accuracy on the Tam et al. NMF labeled dataset to ~97%. This was a substantial improvement over all previous neural network architectures trained on the MAD-5000 gene list. This was considered a reasonable place to stop exploring neural network architecture space, since the remaining misclassified samples had the lowest label confidence from the original NMF assignments. A comparison of the accuracy of the WGCNA based model and the 5000 MAD gene classifier is pictured in Figure 26. Here the subtype assignments from each respective model are plotted along the 2D embedding produced by the autoencoder on the original 5000 MAD gene lists from the original Tam et al. data. Now that I had created a classifier which was accurate at determining subtype, I sought to examine how well this classifier generalized to incoming patient data from the New York Genome Center ALS consortium which was released during the construction of this classifier. This final classifier will be referred to as the WGCNA-NN in this text.

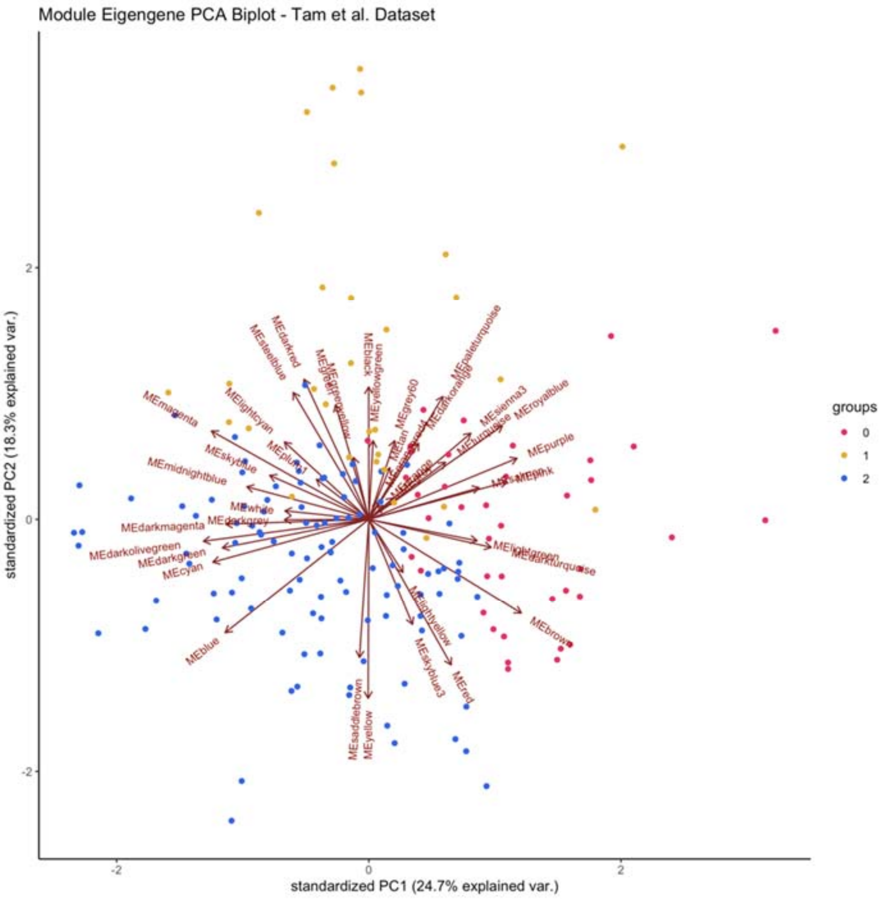


Figure 24 A PCA biplot showing each WGCNA module's contribution to the variance in the Tam et al. dataset

Colors represent the original NMF labels, where 0 is the TE group, 1 is the glial activation group, and 2 is the oxidative stress group. This shows how modules contribute push towards discriminating subtype along the first two principal components.

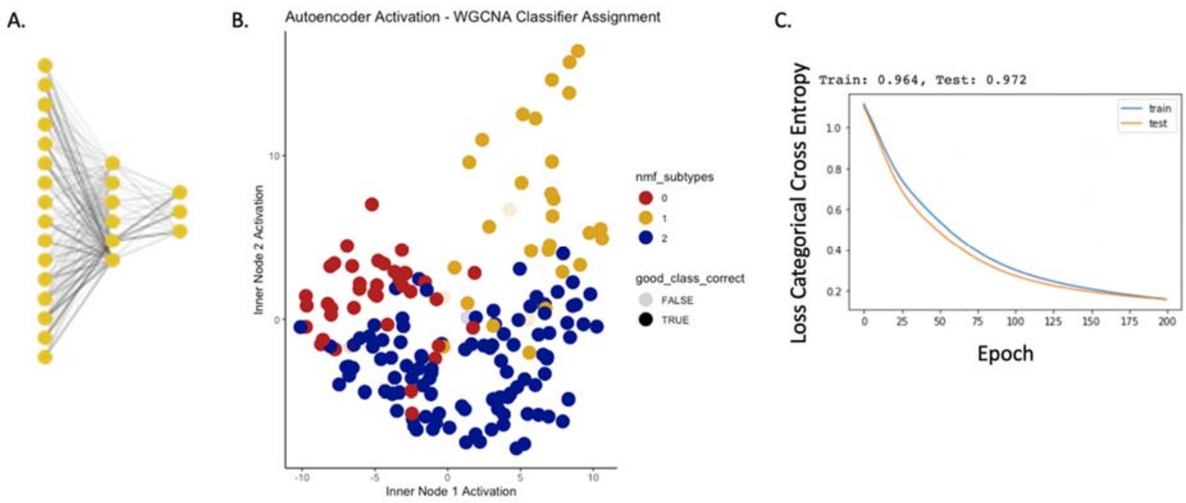


Figure 25 An overview of the WGCNA based classifier

Panel A represents a schematic of the classifier structure with 39 input nodes, 6 inner nodes and 3 output nodes. Panel B depicts the incorrectly called samples in grey on the same autoencoder activation plot featured in Figure 20. Panel C is a plot of loss by categorical cross entropy of the samples in the training and test set after each epoch of training. The final accuracy of predictions on the test and training sets from this classifier are depicted in the title of this panel at 97.2% and 96.4% respectively.

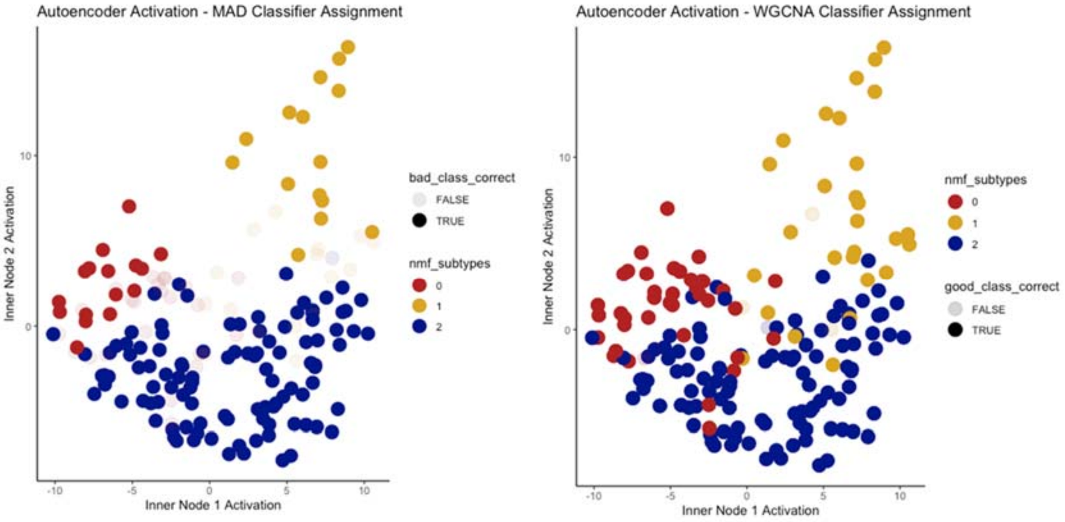


Figure 26 Comparison of Figure 21 and 25

A replotting of panel A from Figure 21 and panel B from Figure 25 for easy comparison and visualization of the improvement between the 5000 MAD and WGCNA based classifiers.

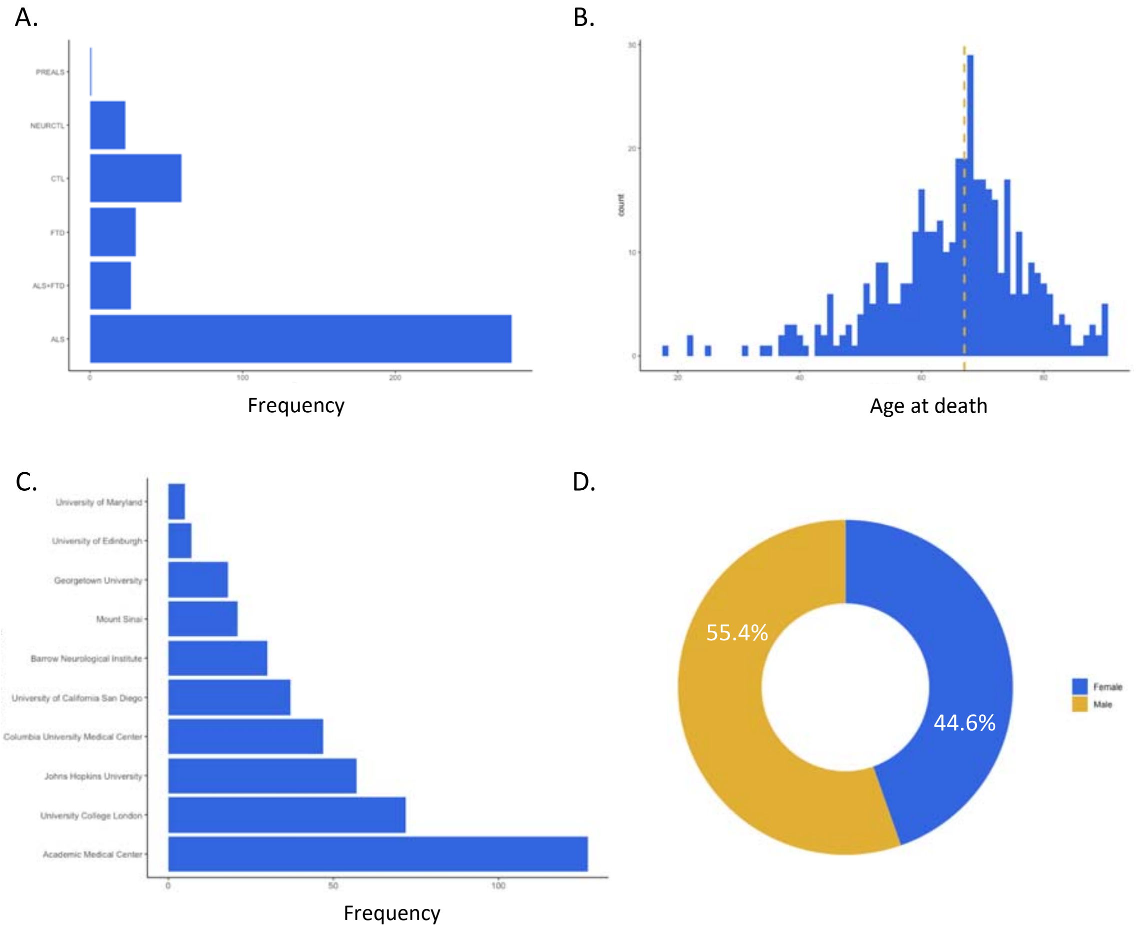


Figure 27 A figure describing basic features of the NYGC and Tam et al. cohort

Panel A depicts the frequency of patients with each disease state, NEURCTL represents controls which have a neurological disorder other than ALS. Panel B is a histogram describing the distribution of ages at death of patient and controls. Panel C is a barplot describing the collection sites of the patient data. Panel D represents the distributions of the sex of the patients.

4.5 Application of a neural network classifier to a novel NYGC ALS and FTD-TDP patient cohort

During the construction of a suitable classification algorithm for ALS subtypes, an additional 543 samples were released by the New York Genome Center (NYGC) ALS Consortium. These samples corresponded to 326 patients and controls constituting a cohort triple the size of the original dataset. The samples included 206 ALS patients, 21 patients with ALS and FTD, 29 patients with FTD, and 67 controls, as shown in Figure 27 Panel A. One additional patient was re-classified as pre-ALS upon detection of an ALS causal gene mutation in their whole genome sequence data (a C9orf72 repeat expansion, known to cause ALS and FTD with high penetrance). An overview of the complete cohort including the Tam et al. data is depicted in Figure 27. These 543 new samples were never used in the training of any neural network for this thesis, they were simply classified using the pre-trained WGCNA-NN classifier. However, the proceeding figures will include all 719 total samples with their original NMF labels to display the relationship of the ALS subtype classifications of the new data with the original training set. To classify the incoming samples, eigengenes were calculated for the incoming samples using the same module structure as assigned to the original cohort, i.e. genes did not change modules. This is a computationally inexpensive step to perform, as calculating the eigengene is simply the first principal component of the genes in a given WGCNA module. As this classifier is designed to turn an unsupervised learning problem into a supervised learning problem, there are no labels on the new dataset with which to check accuracy. Therefore, assessment of how well the classifier performed is associated with how well the structure of subtype was maintained across the introduction of new data. Additionally, NMF was rerun by Oliver Tam on the new NYGC data to see if new subtypes could be detected, and none were found. Figure 28 represents PCA plots of the eigengenes of the Tam et al. dataset and the new data. This figure is encouraging as the clustering within a particular subtype is largely robust and the structure of these groupings relative to each other is also conserved. Figure 29 depicts the push contributed by each of the co-expression modules in the original Tam et al. dataset and in the complete cohort of 719 samples. Figure 29 shows that the module push defined in the original dataset in the first two principal components is conserved in the new NYGC data, as they remain in similar positions with respect to the centers of each respective subtype classification. This suggests that the genes defined by certain co-expression signatures similarly contribute to the assignment of a subtype in the new dataset. This evidence also suggests the structure of the transcriptional landscape in the new data is similar to that of the old data, and these co-expression modules are conserved in how they contribute to the definition of subtype. Figures 28 and 29 allow us to visualize the placement and variation of the new samples with respect to the old dataset and assess how new predictive calls relate with respect to the structure of the original dataset. The largely similar placement of the original samples when compared within a cohort triple the size, the variance in which could be considerably greater than the original dataset, is an encouraging indicator of the robustness of the classification strategy and of ALS subtype itself.

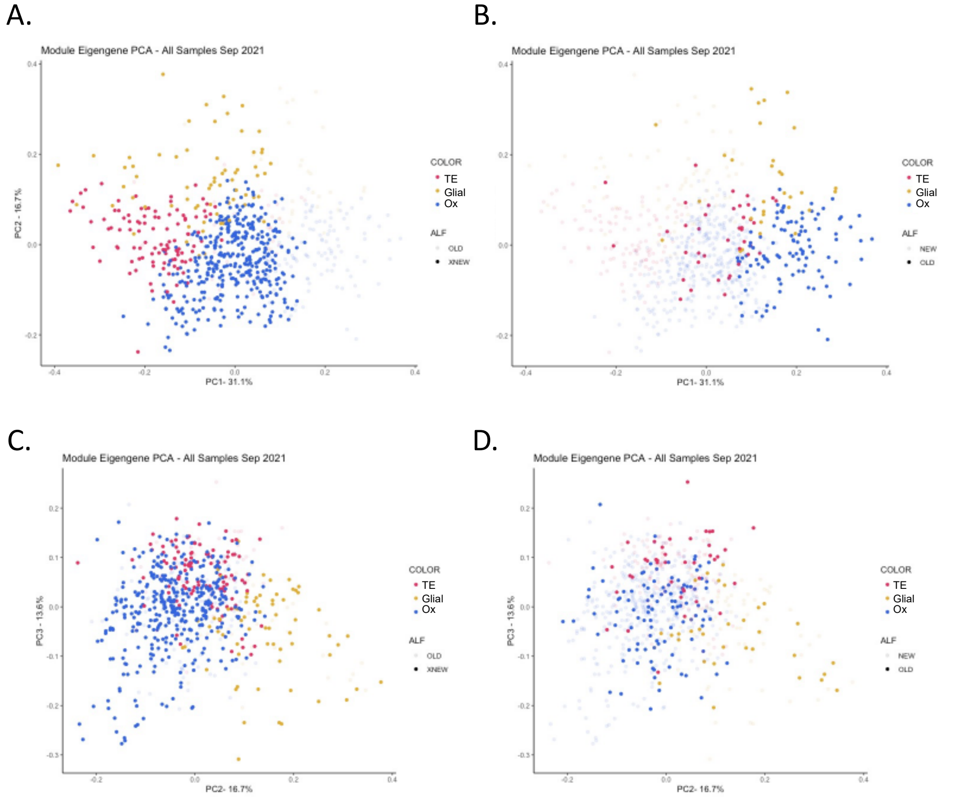


Figure 28 Principal components analysis of the combined Tam et al. and new NYGC ALS consortium cohort

Panels A and B depict principal components 1 and 2 of the calculated module eigengenes, with the samples in the original and novel dataset greyed out respectively. The original NMF labels of the Tam et al. dataset are depicted alongside the novel classifier calls. Panels C and D represent principal components 2 and 3 of the data with the original and novel dataset greyed out respectively. Together these three principal components describe 61.4% of the variance of the data across the two datasets.

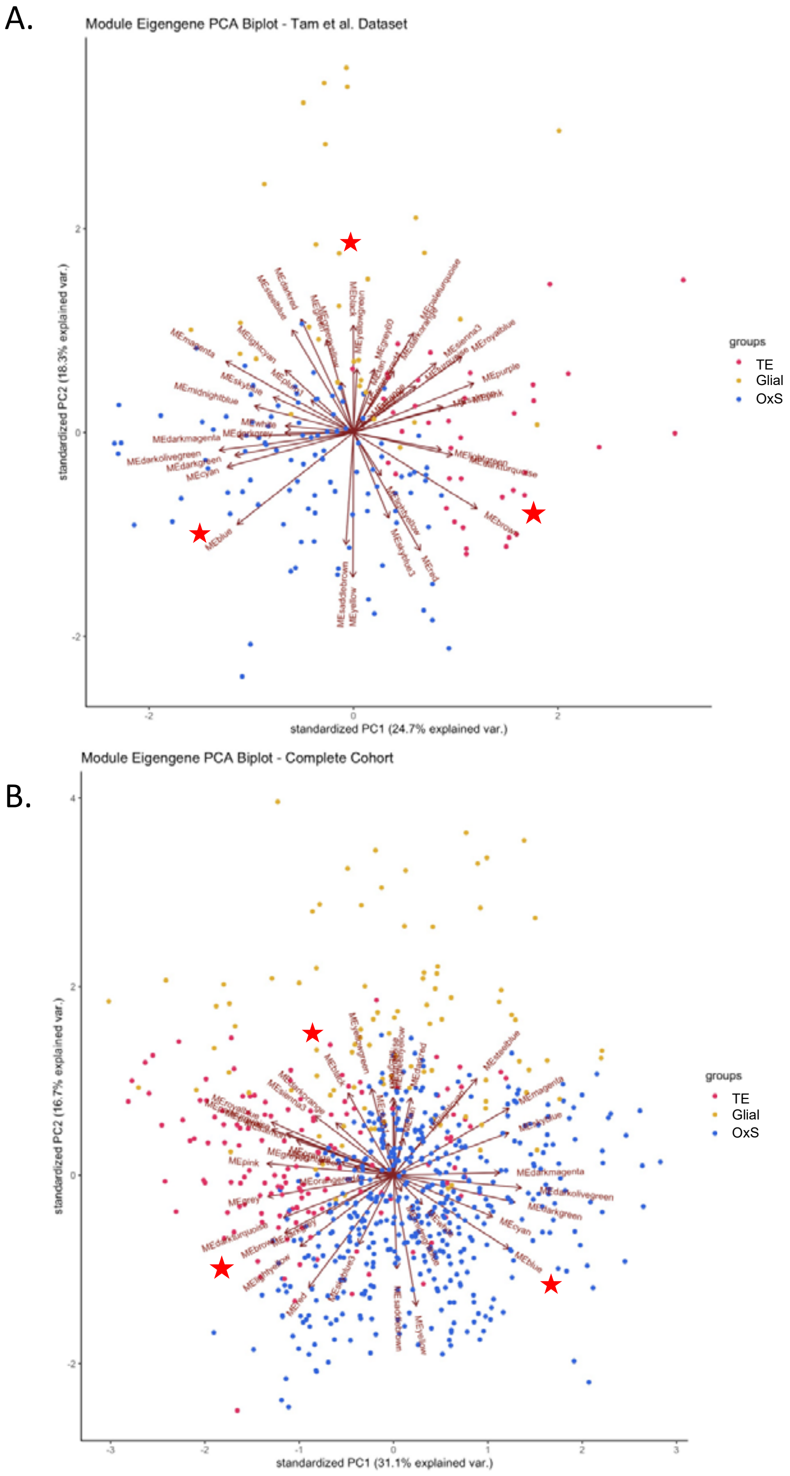


Figure 29 PCA biplots of the push of eigenmodules across the first two principal components of the Tam et al. and novel NYGC patient cohorts

Panel A represents the Tam et al. dataset, and panel B represents the complete cohort.

4.6 Neural net classification expanded to single cell RNA-seq data

In the previous section I described the generalizability of the WGCNA-NN classifier to a large novel patient population. Our lab is interested in exploring whether subtype is present in other data types aside from bulk RNA-seq data, therefore we decided to test the classifier's performance on single cell RNA-seq data obtained from a subset of ALS patients for whom we know the ALS subtype. Three single-cell libraries we collected from tissues of ALS patients and a control in the Tam et al. dataset for which we had bulk-RNA seq NMF classification assignments. However, single cell data is much sparser than bulk RNA-seq data therefore, the original WGCNA eigenmodules used for classification were not good representations of co-expression modules captured within a single cell library. That is, many modules may have been reduced to one or zero genes with detectable expression. However, at this point I had gained confidence in WGCNA as a useful strategy for feature selection and dimensionality reduction, and sought to make a new classifier after having performed WGCNA on the reduced gene set that was produced by single cell libraries. As the description of subtype represents a bulk RNA-seq profile with many cell types, a classification on the profile of each single cell did not seem particularly amenable as a first step. Instead, I used pseudobulked samples as the data for the purposes of classification. A pseudobulked sample is an aggregate formed by the simple summation of counts from all of the cells from a particular sample followed by VST normalization by DEseq2. This pseudobulked sample is still useful for its underlying single cell resolution, because we still have the individual single cell profiles and are able to look at the distribution of cell types within each classified sample. After taking the overlapping gene set of the three pseudobulked test samples with the original bulk-RNA seq data, I recomputed a WGCNA network on these ~5000 genes detectable by both platforms. The user defined optimal power parameter was determined again to be 9 through the same metrics as depicted in Figure 23 computed for this new network seen in Figure 30. The network resulted in 16 co-expression modules depicted in Figure 31. Labels in Figure 31 are the original NMF labels, and plot displays that subtype grouping and structure is conserved across the first two principal components of the eigengenes produced from the reduced single cell gene model of the samples in the Tam et al. dataset. This is encouraging that a trained neural net classifier would be able to discriminate between these subtypes on approximately 25% of the original genes used from the bulk RNA-seq data. To test this, I trained another neural network which took these 16 co-expression modules in as input. I did not have to remove any modules encoding sex as this was not a confound in this reduced set, as displayed in Figure 32. In order to train this new network to detect subtype, I used the eigenvalues of the overlapping gene set from the bulk RNA-seq data in the Tam et al. paper as inputs to the classifier and separated the samples into an 80/20 train test set identical to the split used in the previous sections. The result of this training regime is depicted in Figure 33. Of the three samples we submitted for single cell sequencing, all three predictions using this pseudobulk single cell WGCNA-NN classifier matched the original NMF derived assignments. This data is preliminary, however we are strongly encouraged to collect more single-cell data sets and test the robustness of this pseudobulk single-cell classifier, as we intend to expand this analysis to understand the cellular composition of subtype.

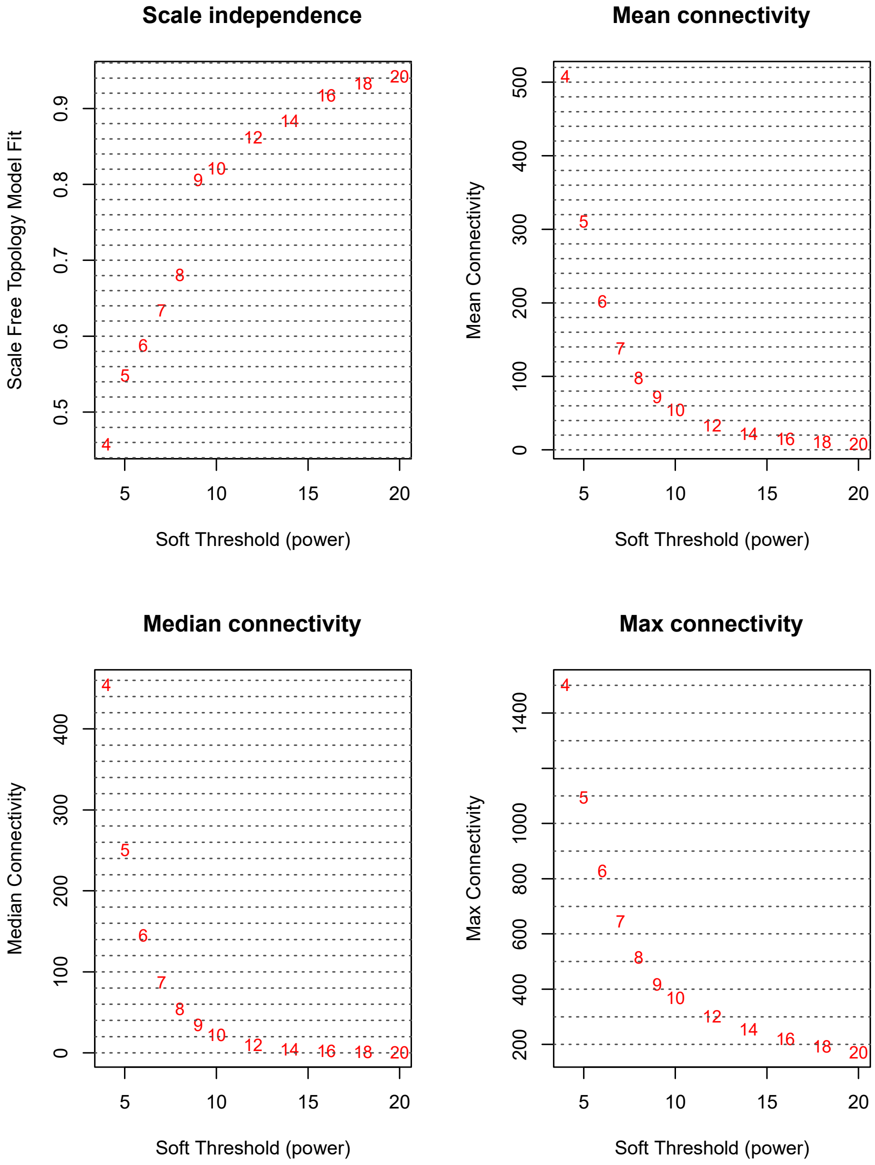


Figure 30 Multiple selection criterion to determine the exponential/power parameter for the decay of the adjacency function in WGCNA for the single cell pseudobulk gene list

The upper left panel plots the the R^2 Pearson correlation between log(p(k)), where p(k) ~ k^{-power} is the connectivity of a particular fit adjacency matrix, and log(k) where k is the connectivity of a scale free network for a set of user defined power values. A value of 1 would indicate the fit network perfectly satisfies the scale free criterion, but in practice a R^2 value of ~.8 is satisfactory. For this analysis we chose a power parameter of 9 as it is near .8 and maintains a higher connectivity within the network. The remaining panels of the figure show the summary statistics of the connectivity of the network for each user defined power parameter. Connectivity is defined as the sum of the connection strength/adjacency of a gene with all other genes in the network.

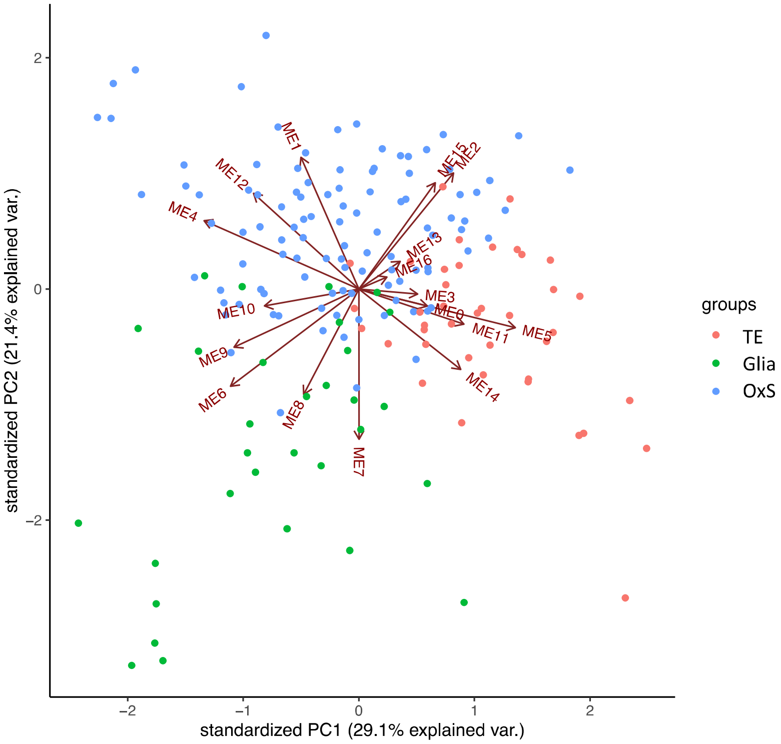


Figure 31 A biplot of the WGCNA derived eigengenes across the first two principal components of the Tam et al. dataset with original NMF labels using a pseudobulk gene list

This plot shows that subtype is disparate even when using a reduced set of 5000 genes.

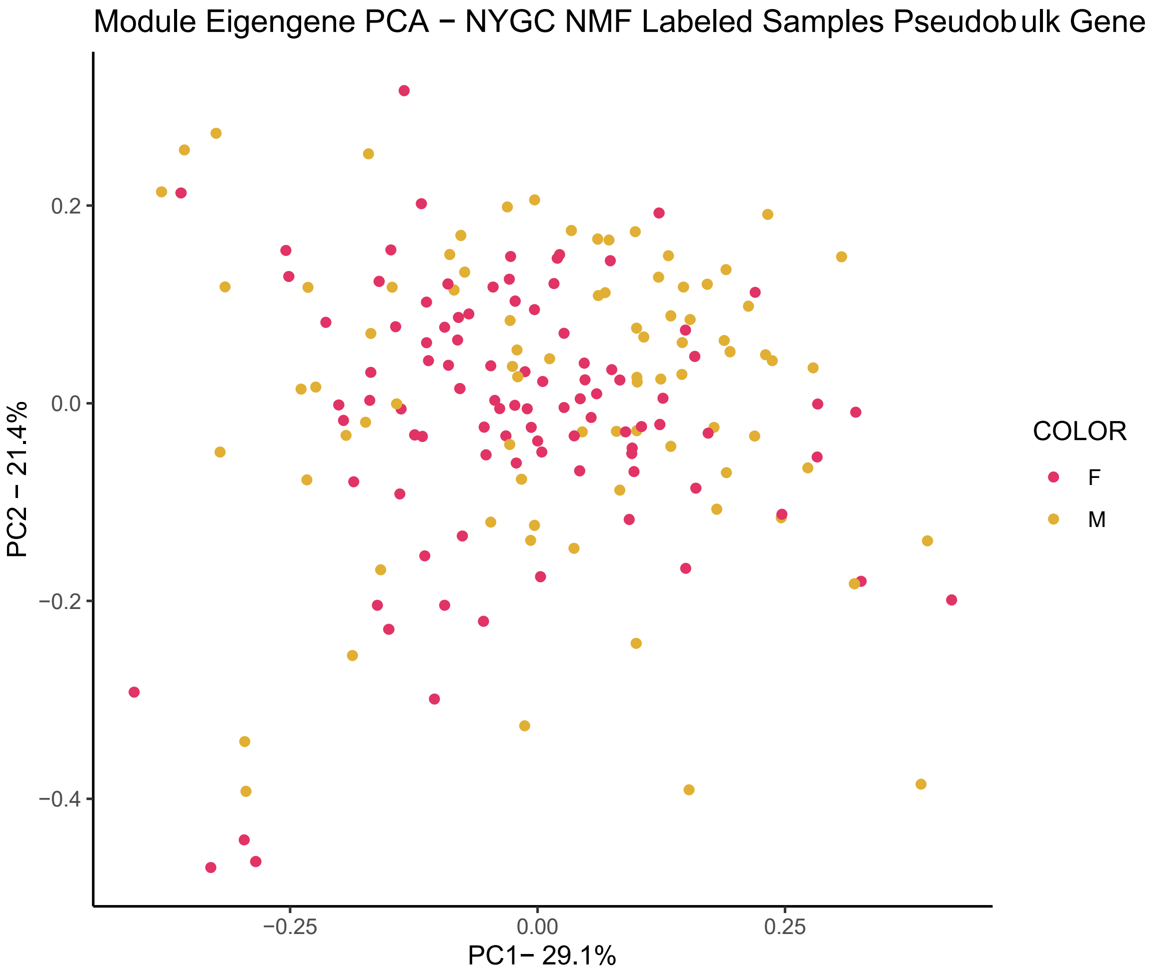


Figure 32 A figure depicting no sex bias across the first principal components of the eigengenes computed from a reduced gene list associated with single cell libraries

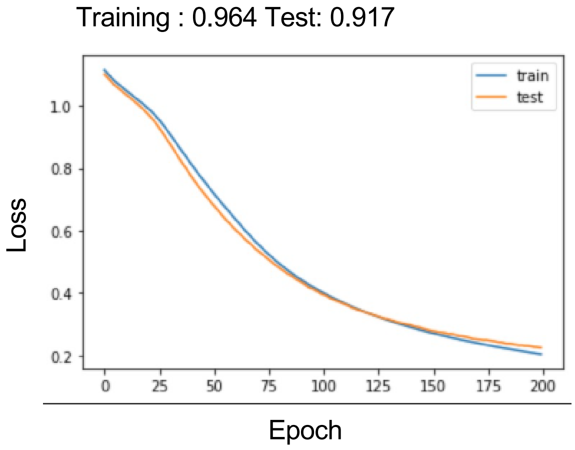


Figure 33 A figure describing the training regime of the single cell pseudobulk WGCNA-NN classifier

This is a plot of loss by categorical cross entropy of the samples in the training and test set after each epoch of training. The final accuracy of predictions on the test and training sets from this classifier are depicted in the title of this panel at 91.7% and 96.4% respectively.

4.7 Conclusion

In this chapter I have described the construction of a neural network classification method by which to easily ascribe a label of ALS transcriptional subtype defined in Tam et al. to a novel patient derived *post mortem* frontal and motor cortex tissue sample. There are three molecular ALS subtypes: a TE activation group (ALS-TE), a microglial activation group (ALS-Glia), and an oxidative stress group (ALS-Ox). I have shown that the structure of these transcriptional subtypes is robust through their detection agnostically by a neural network called an autoencoder. I have shown that this subtype structure is present in a novel dataset three times the size of the data set used to originally characterize ALS subtype. The robustness is underscored by the fact that the same co-expression modules/gene sets as defined by WGCNA separate these disparate subtypes from each other in the original and expanded patient cohorts. Those modules which provide the greatest push towards a particular subtype are an exciting source for investigation. However, this relationship is complicated and will require more extensive analysis than simple GO term enrichment as a module's relationship to a subtype is not 1:1. Understanding how a module with information about multiple subtypes describes a transcriptional dataset will likely require the development of some additional statistical analysis methods outside the scope of this thesis. In addition to creating a classifier for bulk RNA-seq data, I also created a single cell classifier, which from preliminary data appears to be potentially useful for determining how cell type composition contributes to overall ALS subtype. This will be an important link as we determine how changes to cell state within particular cell types contribute to the overall transcriptional profiles present in ALS affected tissues. In attaining the aim of creating a robust classifier, I exemplified how feature selection is an incredibly important step in the development of any machine learning algorithm, and that a technique motivated by the domain of application, for example transcriptome biology, can greatly improve model performance. This work serves to validate that the subtype observed in Tam et al. is a strong biological signal, and has opened up important avenues for investigation as we continue to unravel what subtype means for the etiologies of ALS and FTD-TDP.

Chapter 5: Conclusions and Perspectives

5.1 Overview of conclusions

This thesis has touched upon multiple fields within transposon biology, small RNA biology, computational biology and applied machine learning. I have published a small RNA analysis pipeline which allows for the facile analysis of small RNA libraries and improved this pipeline in several ways after publication. I performed data analysis to understand the nature of a TDP-43 loss of function mutation on TE associated sRNA biology in human somatic cells. I contributed a review of current computational methods tackling the difficult problem of quantifying highly ambiguous TE derived reads in multiple sequencing-based assays. Finally, I contributed several follow up investigations into the nature of ALS/TDP-43-FTD molecular subtype and constructed a classification algorithm with which new incoming patient samples can be classified. This classifier shows decent generalizability to related datasets, and potential to accurately call subtype on single cell sequencing derived data. From these investigations, I conclude that there is significant room for progress on all fronts touched by this thesis, and am enthusiastic for the continuation of the work. In the following sections I will describe my conception of these frontiers and where I would be most interested in seeing development.

5.2 Frontiers of interest in sequencing based analysis

Currently, the majority of RNA and sRNA downstream sequencing analysis is achieved by differential expression analysis, where a case and control are compared to look for expression differences based on a single treatment or condition. This can be expanded to include multiple conditions with the use of generalized linear models in packages like

DEseq2 (Love et al. 2014), however these packages do not address the fundamental problem that all sequencing-based data is measuring relative abundance and is *compositional* data. This means, that our standard differential expression analyses are most accurate when the composition of the sequencing libraries being compared are similar. Still, standard differential expression analysis has proven to be a highly effective approach, particularly when looking for gross changes to gene expression. In some cases, we may not be worried if the composition of our data changes significantly, as these large compositional changes may be precisely what we were interested in. However, if we would like to perform an analysis in which we are looking to detect smaller changes within a particular fraction of a sequencing library, and there are large differences between the abundance of that fraction across libraries, then the bias of our current tools to reporting on the larger compositional changes will present issues. This is an important problem to address as we seek to integrate data from multiple tissues or cell types, and from assay types which are expected to have distinct compositions. This problem is quantified, however not statistically addressed by TEsmall, as it reports the proportions of each of the sRNA categories observed in a library. This is a first step to acknowledging whether gross compositional shifts are a concern in a particular analysis, however, novel computational and experimental methods will need to be developed to formally address this problem. There is evidence to support this as a particular issue for small RNAs, as sRNA composition varies widely across biofluids (Godoy et al. 2018). Comparisons between bulk RNA-seq data from patient tissues with varying levels of neuronal loss in neurodegenerative disease is a another good example, particularly relevant to my thesis, where accurate methods for cellular deconvolution would be useful. It is challenging to look for changes in glial biology between these samples, because one library may contain a significantly higher proportion of glial cells or a more prominent glial transcriptional signature. We can observe gross changes to glial biology, like the elevation of pro-inflammatory microglial markers, but not more subtle changes which might be relevant to the etiology of a sporadic disease like ALS. This is of particular importance when we are striving to elucidate the underlying meaning of a molecular subtype in a disease where cell type compositional changes are inherent in the progression of the disease.

This compositional data challenge, however, is not disease specific. You can also imagine these compositional discrepancies occurring broadly when integrating multiple - omic data types. While inherent compositional discrepancies might not matter initially for characterization across data types, especially if it is the same compositional bias across all data types, I suspect that as the field progresses, we will become more interested in detecting changes within a compositional fraction independently from changes in other fractions of the whole. This can only be accomplished through compositional methods (Aitchison 1982). Attempts to integrate single cell data with bulk samples provides an example of better methods for compositional analysis could be particularly helpful. This type of analysis is important because there are limits to single cell sequencing technology, and it is not always fiscally or technically feasible to collect sufficient numbers and types of single cell data from human patient samples, so bulk RNA-seq is often used. Bulk RNA-seq is especially helpful if a tissue contains cell types which are known to be difficult isolate for capture with single cell sequencing methods like microglia. Additionally, single cell sequencing is known to have a bias toward capturing the most highly expressed transcripts in a cell, though some lowly expressed genes may nevertheless have large impacts on cellular function. Looking at these highly expressed transcripts we are able to largely determine the cell type of a particular cell using correlation with reference atlases from consortia that have spent great effort in generating atlases of characterized cell types. However, it is still difficult to robustly assess subtle differences in cellular states within a given cell type, especially if the disease relevant genes are in the lowly expressed repertoire of the cell. Using a mixture of bulk and single-cell data profiles can be helpful in allowing one to characterize the composition of a library from the highly expressed genes detected in single cells, and the more lowly expressed genes across that sample from the bulk RNA-seq library. The big challenge then becomes how to deconvolve these two relative measurements so that the researcher can benefit from the merit of the specificity of single-cell RNA-seq and the accessibility and depth of bulk RNA-seq. One example of how this is currently done involves pseudobulking: aggregating the expression from all cells from a single-cell library and then comparing that with a matched sample from a bulk RNA-seq library. It it is unclear whether this can be used to accurately deconvolve the signals from bulk expression data based down to a single-cell level based on the single cell atlas derived labels. However, this is an excellent first step to look at the type of discrepancies we may encounter while attempting to match compositional biases with current methods.

The idyllic tool I would create to address this problem would be one which could match the high abundance repertoire of the single cell library and robust cell type labels, to the signal of a cell's particular high abundance repertoire and its proportion of expression in the bulk RNA-seq library. One glaring problem is where is the proportional reference located in a sequencing library, a proportional fold change with respect to what cell type or disease state for example, and how does one integrate differences in reference points across assays and data types to glean meaningful insights? The reference cannot be a cell expressing nothing, because one cannot compute a proportional change away from zero, but a flat expression profile of a cell might also create large biases when estimating the proportion of counts between cells or some level of disease state because it is artificial. An approach could be to create a non-informative cell, or a cell expression profile which does not contain any information, and resides at the center of the simplex of the compositional data a researcher is trying to integrate. This may help us begin to ground how changes in expression proportions push toward one cell type, condition, or other variable which would affect composition of the repertoire, but does not help in the cases where a gene is absent from one condition to another. Integration of zero counts into proportional data is an active field of research, and considering genomic data is proportional data consisting largely of zeros is a glaring problem (Greenacre 2021). This question of reference in compositional data is fundamental to -omic data integration and currently we are circumnavigating this problem through aligning lower dimensional embeddings which capture shared gross compositional changes across data points and datasets. I employed this method with some success in this thesis. However, it is not yet clear to me whether the alignments of these embeddings is not at least somewhat arbitrary. Moreover, these embeddings rely heavily on robust generalities of compositional changes between conditions. I suspect that as we refine our understanding of biology through -omic data we will, as we often do, encounter more exceptions than rules. Therefore, robust compositional analysis will be a fruitful path of investigation.

5.3 Paths of development from ALS/FTD-TDP cohort analysis

The latter part of this thesis has been dedicated to the analysis of clinical data from ALS and FTD-TDP-43 patients and controls. While this work was beneficial as an early step toward the characterization of patient subtype, there is significant room for growth in our understanding of these subtypes in ALS. Researchers have noted previously that some of the heterogeneity in ALS involves multiple independent contributing phenomena, including but not limited to, blood brain barrier (BBB) compromise and TDP-43 aggregate pathology (Waters et al. 2021). However, how these multiple factors combine to give rise to a postmortem subtype assignment is largely unknown. For this particular example, there is some indication that our described ALS subtype associated with microglial activation was also associated with blood brain barrier compromise, as I observed increased expression of genes associated with the BBB is this group. Understanding the role of multiple pathological phenomena through their coalescence in a single time point at end stage is a very difficult problem, and the fact that we see a robust signature of subtype in a large cohort of patients is alone a striking feature in our data. This would indicate that there are a finite number of strong and robust endpoints at which many patients with ALS will arrive, even if we are currently unsure as to what those arrival points represent with respect to the etiology and pathophysiology of ALS.

The robustness of subtype is encouraging as it may help us in the development of diagnostic technologies. Of particular interest to the field is the presence of biomarkers, or substrates which can be readily assayed from peripheral tissues or biofluids that indicate the status of a disease. Because subtype is robust, it is of great interest to link biomarkers with subtype as potential diagnostic aids. This signal could be potentially found as a direct reporter of subtype in the patient, or recovered through the differentiation of patient derived fibroblasts into induced motor neurons (iMNs) and calling subtype with a classifier directly in this system. Both approaches have obvious caveats. Subtype as it is currently defined is the result of some progression of neurodegenerative factors as their transcriptional landscape presents in the frontal and motor cortices. It is probable, but certainly not assured, that some eminence of subtype will be present in related tissues, like blood, cerebrospinal fluid, or muscle. Currently, our lab is interested in collecting matched tissue with frontal and motor cortical biopsies to create a dataset of matched peripheral tissue and cortex, with the intention of calling subtype in cortex and searching for correlated transcriptional markers in other associated tissues. In tandem with this objective, we are looking in other tissue databanks like Answer ALS, where there are repositories of induced motor neurons from patient derived fibroblasts. This approach of looking for subtype in iMNs relies heavily on the fact that although ALS is a largely sporadic disease, we anticipate there is underlying genetic variation, combined with environmental factors, which predisposes an individual toward a particular state in ALS. Obviously, the brain tissue is subject to a variety of other environmental factors that mesodermal skin cells are not, but this is the current frontier of the field however imperfect. One short term approach is to collect richer patient histories to robustly annotate and score additional factors, for example head injuries or pharmacological treatment for mental illness, or diseases which may cause systemic inflammation especially in the gut. I have been able to call subtype, with reasonably high confidence in patient derived iMNs, however as subtype was originally characterized from bulk RNA-seq data which contains a multiplicity of cell types, it is uncertain how much signal contributed by non-neuronal cells was lost in these iMN specific calls. This relates to the problem of integration of compositional data discrepancies across multiple data types. As I described earlier in this chapter, if one was to detect a signal associated with the glial subtype in iMNs, how would one begin to detect this signal in the cortical derived samples classified into the glial group, where the neuron derived signal is potentially weaker compared to that of the other components coming from glial cells? Hopefully, further investigations into biomarkers will be fruitful, and allow for the stratification of ALS patients, as it may provide a path to select candidates for which clinical trials may be more effective. This would be an immediate step that could be helpful even as we as a field continue to unravel the etiology of the incredibly complicated landscape of the neurodegenerative disease ALS.

I have spent the majority of this thesis discussing physical properties of biological phenomena and their potential role in disease; however, no system exists in isolation. Prolonged investigation into the mechanisms by which a nervous system deteriorates has led me to deeply consider the environment in which a mind thrives, and that perhaps our scope as biologists is too limited considering only the physiological phenomena at the root of disease. A neurodegenerative diagnosis is a fast-paced progression towards death in which we lose our grasp on one of the most precious resources we have as human beings, our contact with the home that is mind and/or our body. During the 2021 Chan Zuckerberg Initiative conference for neurodegeneration, there was a breakout session in which patients were asked how we could help as researchers, and one African-American mother replied that she wished there were more resources to support the mental and emotional wellness of their family as they confronted the staggering diagnosis of her daughter. As a biologist, I would be remiss if I did not include some commentary upon the context in which this work was conducted, as we well know context is everything when attempting to understand life. As I write this, we currently exist amongst a menagerie of concurrent crises. We are collectively experiencing a global pandemic which has killed over 6 million people worldwide and with the inadequate and unaffordable healthcare provided in the United States, is likely to profoundly exacerbate the already gaping wealth inequality in America. This wealth gap disparately affects people of color who on average possess a small fraction of the generational wealth of white Americans (Lettieri 2021). According to the 2019 survey of consumer finances by the Federal Reserve the average net worth of the bottom 25% of

American families is $310, and the bottom 50% of families possess only 2% of the total of US wealth (Lettieri 2021). This financial insecurity has placed extraordinary pressure on individuals, and has many healthcare workers on track to not be able to afford the services they provide. This in my mind, recontextualizes my work as a biologist, as any merits derived from my research toward human disease are likely to be largely inaccessible to the majority of patients in America. For this reason, it is important for me to consider adapting and expanding my expertise into fields with the potential to address the health and wellness of individuals in addition to the heroic interventions of traditional western medicine. I also think it is of particular import to investigate how racial and socioeconomic pressure affects health outcomes, the effects that the stress from these pressures has on human biology, and if there are low/no-cost interventions which can help mitigate the effects of this chronic stress and trauma on the human body. It has been shown that increased maternal stress and cortisol levels are associated with differential methylation of imprinted genes in the child (Vangeel et al. 2015), and that chronic or acute maternal stress associated with war can affect genes regulating the hypothalamic-pituitary-adrenocortical system in the child (Kertes et al. 2016). I believe that this is an excellent arena in which to apply my understanding of transposon biology, genome biology, and neurological disease, as I aspire to develop methods to improve health intergenerationally and reduce inequities.

5.4 Positionality

I am a fair-skinned Brown woman of mixed African-American and white descent. I am deeply proud and with a joyful heart assert my Blackness. I am the first member of my immediate family to attain a four-year degree, and come from a lower/lower-middle class family. I attended community college before transferring to a four-year university where I was a McNair Scholar and Regent’s Scholar. I speak English as my first language, and spent between two and three of my formative years living abroad in Mexico and in Chile where I attended school and learned Spanish. I am a scientist and a seeker.

Chapter 6: References

Abadi M, Agarwal A, Barham, Paul, Brevd E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/.

Adusumalli S, Feroz Mohd Omar M, Soong R, Benoukraf T. 2014. Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief Bioinform* **16**: 369–379.

Aitchison J. 1982. The Statistical Analysis of Compositional Data. *J R Stat Soc Ser B* **44**: 139–160.

Alcala AM, Flaherty KT. 2012. BRAF Inhibitors for the Treatment of Metastatic Melanoma : Clinical Trials and Mechanisms of Resistance. *Clin Cancer Res* **18**: 33–40.

Allshire RC, Madhani HD. 2018. Ten principles of heterochromatin formation and function. *Nat Rev Mol Cell Biol* **19**: 229–244. http://dx.doi.org/10.1038/nrm.2017.119.

Amselem J, Cornut G, Choisne N, Alaux M, Alfama-Depauw F, Jamilloux V, Maumus F, Letellier T, Luyten I, Pommier C, et al. 2019. RepetDB : a unified resource for transposable element references. *Mob DNA* **10**: 4–11.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: 1–12.

Anders S, Pyl PT, Huber W. 2015. HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Mol Cell* **31**: 785–799.

Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, Zagalak JA, Emmett W, Smith CWJ, Luscombe NM, et al. 2018. Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**: 1067–1081.

Axtell MJ. 2014. Butter: High-precision genomic alignment of small RNA-seq data. *bioRxiv* 1–16.

Axtell MJ. 2013. ShortStack : Comprehensive annotation and quantification of small RNA genes ShortStack : Comprehensive annotation and quantification of small RNA genes. 740–751.

Babaian A, Thompson IR, Lever J, Gagnier L, Karimi MM, Mager DL. 2019. LIONS: analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *Bioinformatics* 1–3.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update , a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 4–9.

Bartel DP. 2018. Metazoan MicroRNAs. *Cell* **173**: 20–51. http://dx.doi.org/10.1016/j.cell.2018.03.006.

Baruzzo G, Hayer KE, Kim EJ, Camillo B Di, Fitzgerald GA, Grant GR. 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat Methods* **14**: 135–139.

Bendall ML, Mulder M De, Iñiguez LP, Lecanda-Sánchez A, Pérez-Losada M, Ostrowski MA, Jones RB, Mulder LCF, Reyes-Terán G, Crandall KA, et al. 2019. Telescope : Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLOS Comput Biol* 1–25.

Bennett EA, Keller H, Mills RE, Schmidt S, Moran J V., Weichenrieder O, Devine SE. 2008. Active Alu retrotransposons in the human genome. *Genome Res* **18**: 1875–1883.

Berger S, Pachkov M, Arnold P, Omidi S, Kelley N, Salatino S, Nimwegen E van. 2019. Crunch: Integrated processing and modeling of ChIP-seq data in terms of regulatory motifs . *Genome Res* 1–24.

Bishop CM. 2006. *Pattern Recognition and Machine Learning*. 1st ed. Springer New York.

Boissinot S, Chevret P, Furano A V. 2000. L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History. *Mol Biol Evol* **17**:

915–928.

Bokeh Development Team. 2014. Bokeh: Python library for interactive visualization. http://www.bokeh.pydata.org.

Boroviak T, Stirparo GG, Dietmann S, Hernando-Herraez I, Mohammed H, Reik W, Smith A, Sasaki E, Nichols J, Bertone P. 2018. Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development* **145**: 1–18.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biol* **19**: 1–12.

Bourque G, Leong B, Vega VB, Chen X, Yen LL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Huck HN, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**: 1752–1762.

Bousios A, Gaut BS, Darzentas N. 2017. Considerations and complications of mapping small RNA high-throughput data to transposable elements. *Mob DNA* **8**: 1–13.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Brocks D, Chomsky E, Mukamel Z, Lifshitz A, Tanay A. 2018. Single cell analysis reveals dynamics of transposable element transcription following epigenetic de-repression. *bioRxiv*.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran J V, Kazazian HH. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci* **100**: 5280–5285.

Buenrostro J, Wu B, Chang H, Greenleaf W. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**: 1–10.

Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, et al. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (80-)* **357**: 661–667.

Cao Y, Chen G, Wu G, Zhang X, McDermott J, Chen X, Xu C, Jiang Q, Chen Z, Zeng Y, et al. 2019. Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res* **29**: 1–13.

Cao Y, Chen G, Wu G, Zhang X, McDermott J, Chen X, Xu C, Jiang Q, Chen Z, Zeng Y, et al. 2018. Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res*

40–52.

Carnevali D, Conti A, Pellegrini M, Dieci G. 2017. Whole-genome expression analysis of mammalian-wide interspersed repeat elements in human cell lines. *Dna Res* **24**: 59–69.

Castañeda J, Genzor P, Bortvin A. 2011. piRNAs, transposon silencing, and germline genome integrity. *Mutat Res - Fundam Mol Mech Mutagen* **714**: 95–104.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1–16.

Chollet F, Others. 2015. Keras. https://github.com/fchollet/keras.

Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey CN, Keles S. 2011. Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data. *PLoS Comput Biol* **7**.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**: 71–86.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (80-)* **351**: 1083–1088.

Claycomb JM. 2014. Ancient endo-siRNA pathways reveal new tricks. *Curr Biol* **24**: R703–R715. http://dx.doi.org/10.1016/j.cub.2014.06.009.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**: 215–219.

Cole C, Byrne A, Beaudin AE, Forsberg EC, Vollmers C. 2018. Tn5Prime , a Tn5 based 5′ capture method for single cell RNA-seq. *Nucleic Acids Res* **46**: 1–12.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**: 691–703.

Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. 2014. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**: 1–17.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*

27: 3423–3424.

Daron J, Slotkin RK. 2017. EpiTEome : Simultaneous detection of transposable element insertion sites and their DNA methylation levels. *Genome Biol* **18**: 1–10.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet* **7**.

Dehkordi SK, Walker J, Sah E, Bennett E, Atrian F, Frost B, Woost B, Bennett RE, Orr TC, Zhou Y, et al. 2021. Profiling senescent cells in human brains reveals neurons with CDKN2D/p19 and tau neuropathology. *Nat Aging* **1**: 1107–1116.

Déléris A, Berger F, Duharcourt S. 2021. Role of Polycomb in the control of transposable elements. *Trends Genet* **37**: 882–889.

Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. *Nat Rev Genet* **20**.

Deschamps-Francoeur G, Boivin V, Sherif AE, Scott MS. 2019. CoCo: RNA-seq read assignment correction for nested genes and multimapped reads. *Bioinformatics* 1–9.

Díaz-Martínez M, Benito-Jardon L, Alonso L, Koetz-Ploch L, Hernando E, Teixido J. 2018. miR-204-5p and miR-211-5p contribute to BRAF inhibitor resistance in melanoma. *Cancer Res* **78**: 1017–1030.

Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT, et al. 2019. Systematic comparative analysis of single cell RNA-sequencing methods. *bioRxiv*.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Edwards JR, Yarychkivska O, Boulard M, Bestor TH. 2017. DNA methylation and DNA methyltransferases. *Epigenetics and Chromatin* **10**: 1–10.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397–405.

Feschotte C, Pritham EJ. 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu Rev Genet* **41**: 331–368.

Fletcher SJ, Boden M, Mitter N, Carroll BJ. 2018. SCRAM : a pipeline for fast index-free small RNA read alignment and visualization. *Bioinformatics* **34**: 2670–2672.

Friedländer MR, MacKowiak SD, Li N, Chen W, Rajewsky N. 2012. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**: 37–52.

Friedli M, Trono D. 2015. The Developmental Control of Transposable Elements and the Evolution of Higher Species. *Annu Rev Cell Dev Biol* **31**: 429–451.

Gatenby R, Brown J. 2018. The Evolution and Ecology of Resistance in Cancer Therapy. *Cold Spring Harb Perspect Med* **8**: 1–12.

Gázquez-Gutiérrez A, Witteveldt J, Heras SR, Macias S. 2021. Sensing of transposable elements by the antiviral innate immune system. *Rna* **27**: 735–752.

Gebert D, Hewel C, Rosenkranz D. 2017. Unitas: The universal tool for annotation of small RNAs. *BMC Genomics* **18**: 1–14.

Ghasemi M, Brown Jr. RH. 2018. Genetics of amyotrophic lateral sclerosis. *Cold Spring Harb Perspect Med* **8**: 37–44.

Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z, et al. 2008. Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. *Science (80-)* **320**: 1077–1081.

Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, Stoye J, Tristem M, Johnson WE. 2018. Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology* **15**: 1–11.

Godoy PM, Bhakta NR, Barczak AJ, Cakmak H, Fisher S, MacKenzie TC, Patel T, Price RW, Smith JF, Woodruff PG, et al. 2018. Large Differences in

Small RNA Composition Between Human Biofluids. *Cell Rep* **25**: 1346–1358. https://doi.org/10.1016/j.celrep.2018.10.014.

Goerner-potvin P, Bourque G. Computational tools to unmask transposable elements. *Nat Rev Genet*. https://www.nature.com/articles/s41576-018-0050-x.

Göke J, Lu X, Chan Y, Ng H, Ly L-H, Sachs F, Szczerbinska I. 2015. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell* **16**: 135–141.

Golan T, Messer AR, Amitai-Lange A, Melamed Z, Ohana R, Bell RE, Kapitansky O, Lerman G, Greenberger S, Khaled M, et al. 2015. Interactions of Melanoma Cells with Distal Keratinocytes Trigger Metastasis via Notch Signaling Inhibition of MITF. *Mol Cell* **59**: 664–676. http://dx.doi.org/10.1016/j.molcel.2015.06.028.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science (80-)* **352**.

Graña O, López-Fernández H, Fdez-Riverola F, Pisano DG, Glez-Peña D. 2018. Bicycle : a bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics* **34**: 1414–1415.

Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA. 2011. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**: 2518–2528.

Greenacre M. 2021. Compositional Data Analysis. *Annu Rev Stat Its Appl* **8**: 271–299. https://doi.org/10.1146/annurev-statistics-042720-124436.

Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, Sammeth M. 2012. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res* **40**: 10073–10083.

Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**: 221–225.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Res* **36**: 70–74.

Guffanti G, Bartlett A, Klengel T, Klengel C, Hunter R, Glinsky G, Macciardi F. 2018. Novel Bioinformatics Approach Identifies Transcriptional Profiles of Lineage-Specific Transposable Elements at Distinct Loci in the Human Dorsolateral Prefrontal Cortex. *Mol Biol Evol* **35**: 2435–2453.

Gunady MK, Mount SM, Bravo HC. 2018. Fast and interpretable alternative splicing and differential gene-level expression analysis using transcriptome segmentation with Yanagi. *bioRxiv* 1–23.

Guo J, Grow EJ, Mlcochova H, Maher GJ, Lindskog C, Nie X, Guo Y, Takei Y, Yun J, Cai L, et al. 2018. The adult human testis transcriptional cell atlas. *Cell Res*.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.

Hadi LHA, Lin QXXL, Minh TT, Loh M, Ng HK, Salim A, Soong R, Benoukraf T. 2018. miREM: an expectation-maximization approach for prioritizing miRNAs associated with gene-set. *BMC Bioinformatics* **19**: 1–8.

Haghshenas E, Sahinalp SC, Hach F. 2019. lordFAST : sensitive and Fast Alignment Search Tool for LOng noisy Read sequencing Data. *Bioinformatics* **35**: 20–27.

Hakim ST, Alsayari M, McLean DC, Saleem S, Addanki KC, Aggarwal M, Mahalingam K, Bagasra O. 2008. A large number of the human microRNAs target lentiviruses, retroviruses, and endogenous retroviruses. *Biochem Biophys Res Commun* **369**: 357–362.

Han BW, Wang W, Zamore PD, Weng Z. 2015. PiPipes: A set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome-and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics* **31**: 593–595.

Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**. http://dx.doi.org/10.1186/s13100-016-0065-9.

Handzlik JE, Tastsoglou S, Vlachos IS, Hatzigeorgiou AG. 2019. Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data. *bioRxiv*.

Hardiman O, Al-Chalabi A, Chio A, Corr EM, Logroscino G, Robberecht W, Shaw PJ, Simmons Z, Van Den Berg LH. 2017. Amyotrophic lateral sclerosis. *Nat Rev Dis Prim* **3**.

Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, et al. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28**.

Hashimshony T, Senderovich N, Avital G, Klochendler A, Leeuw Y De, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, et al. 2016. CEL-Seq2 : sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol* **17**: 1–7.

Ho YJ, Anaparthy N, Molik D, Mathew G, Aicher T, Patel A, Hicks J, Hammell MG. 2018. Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome Res* **28**: 1353–1363.

Hodis E, Watson IR, Kryukov G V, Arold ST, Imielinski M, Theurillat J, Nickerson E, Auclair D, Li L, Place C, et al. 2012. A Landscape of Driver Mutations in Melanoma. *Cell* **150**: 251–263.

Huang KYY, Huang Y-J, Chen P-Y. 2018. BS-Seeker3 : ultrafast pipeline for bisulfite sequencing. *BMC Bioinformatics* **19**: 2–5.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**: 81–89.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**: 90–95.

Hyun K, Jeon J, Park K, Kim J. 2017. Writing, erasing and reading histone lysine methylations. *Exp Mol Med* **49**.

Imbeault M, Helleboid P-Y, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**: 550–554.

International Rice Genome Sequencing Project, Schnable PS, Ware D, Fulton

RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, et al. 2005. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Nature* **326**: 1112–1115. http://www.sciencemag.org/cgi/doi/10.1126/science.1178534%5Cnhttp://www.nature.com/doifinder/10.1038/nature03895.

Islam S, Kjällquist U, Moliner A, Zajac P, Fan J, Lönnerberg P, Linnarsson S. 2012. Highly multiplexed and strand-specific single-cell RNA 5 ′ end sequencing. *Nat Protoc* **7**: 813–828.

Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. 2017. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet* **49**: 1502–1510.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel K V, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36**: 321–323.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Franziska P, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, et al. 2014. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science (80-)* **343**: 776–779.

Jeong H-H, Yalamanchili HK, Guo C, Shulman JM, Liu Z. 2018. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. In *Pacific Symposium on Biocomputing*, pp. 168–179.

Jin Y, Tam OH, Paniagua E, Hammell M. 2015. TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**: 3593–3599.

Jo M, Lee S, Jeon YM, Kim S, Kwon Y, Kim HJ. 2020. The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Exp Mol Med* **52**: 1652–1662. http://dx.doi.org/10.1038/s12276-020-00513-7.

Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol* **17**: 355–370. http://dx.doi.org/10.1038/s41579-019-0189-2.

Jones E, Oliphant T, Peterson P, others. {SciPy}: Open source scientific tools for {Python}. http://www.scipy.org/.

Kahles A, Behr J, Rätsch G. 2016. MMR : a tool for read multi-mapper resolution. *Bioinformatics* **32**: 770–772.

Karlsson K, Lönnerberg P, Linnarsson S. 2017. Alternative TSSs are co-regulated in single cells in the mouse brain. *Mol Syst Biol* **13**: 1–10.

Kawahara Y, Mieda-sato A. 2012. TDP-43 promotes microRNA biogenesis as a component of the Drosha and Dicer complexes. *Proc Natl Acad Sci* **2012**: 1–

6.

Kelley DR, Hendrickson DG, Tenen D, Rinn JL. 2014. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol* **15**: 1–16.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Keon M, Musrie B, Dinger M, Brennan SE, Santos J, Saksena NK. 2021. Destination Amyotrophic Lateral Sclerosis. *Front Neurol* **12**: 1–15.

Kerpedjiev P, Hammer S, Hofacker IL. 2015. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**: 3377–3379.

Kertes DA, Kamin HS, Hughes DA, Rodney NC, Bhatt S, Mulligan CJ. 2016. Prenatal Maternal Stress Predicts Methylation of Genes Regulating the Hypothalamic–Pituitary–Adrenocortical System in Mothers and Newborns in the Democratic Republic of Congo. *Child Dev* **87**: 61–72.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

Koo PK, Eddy SR. 2019. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput Biol* **15**: 1–17.

http://dx.doi.org/10.1371/journal.pcbi.1007560.

Kramer MA. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AlChE* **37**: 233–243.

Krizanovic K, Echchiki A, Roux J, Sikic M. 2018. Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**: 748–754.

Krueger F, Andrews SR. 2011. Bismark : a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.

Krug L, Chatterjee N, Borges-Monroy R, Hearn S, Liao WW, Morrill K, Prazak L, Rozhkov N, Theodorou D, Hammell M, et al. 2017. *Retrotransposon activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS*.

Kruse K, Díaz N, Enriquez-Gasca R, Gaume X, Torres-Padilla M-E, Vaquerizas JM. 2019. Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *bioRxiv* 1–28.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, Fitzhugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–932.

Langfelder P, Horvath S. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Curr Protoc*

Bioinforma 1–24.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* **9**: 1–10.

Lee DD, Seung SH. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788–791. https://www.nature.com/articles/44565.pdf.

Leite K, Morais D, Reis S, Viana N, Moura C, Florez M, Silva I, Dip N, Srougi M. 2013. MicroRNA 100: a context dependent miRNA in prostate cancer. *Clinics* **68**: 797–802. http://clinics.org.br/article.php?id=1091.

Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. 2017. TETOOLS facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res* **45**: 1–12.

Lettieri JW. 2021. *The Income and Wealth Inequality Crisis in America – Testimony before The United States Senate Committee on the Budget*. https://eig.org/news/the-income-and-wealth-inequality-crisis-in-america-testimony#:~:text=The numbers are startling%3A The,families is a mere %24310.&text=The bottom 50 percent of,percent of total U.S. wealth.

Levin HL, Moran J V. 2011. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* **12**: 615–627.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**: 493–500.

Li B, Tambe A, Aviran S, Pachter L. 2017. PROBer Provides a General Toolkit for Analyzing Sequencing-Based Toeprinting Assays. *Cell Syst* **4**: 568–574.

Li D, Luo L, Zhang W, Liu F, Luo F. 2016. A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics* **17**: 1–11.

Li H. 2018. Sequence analysis Minimap2 : pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows – Wheeler transform. *Bioinformatics* **26**: 589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li W, Jin Y, Prazak L, Hammell M, Dubnau J. 2012. Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. *PLoS One* **7**: 1–10.

Li W, Lee MH, Henderson L, Tyagi R, Bachani M, Steiner J, Campanac E, Hoffman DA, Geldern G Von, Johnson K, et al. 2015. Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* **7**.

Lieberman-Aiden E, Berkum NL Van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-)* **326**: 289–294.

Lin H, Hsu W. 2017. Kart: a divide-and-conquer algorithm for NGS read alignment. *Bioinformatics* **33**: 2281–2287.

Lister R, Malley RCO, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**: 523–536.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–322.

Liu B, Gao Y, Wang Y. 2017. LAMSA: fast split read alignment with long approximate matches. *Bioinformatics* **33**: 192–201.

Liu B, Guan D, Teng M, Wang Y. 2015. rHAT : fast alignment of noisy long reads with regional hashing. *Bioinformatics* **32**: 1625–1631.

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, et al. 2013. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**: 1434–1442.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 1–21.

Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, Ng H-H. 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol* **21**: 423–425.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**: 1202–1214.

Mak J, Kleiman L. 1997. Primer tRNAs for reverse transcription. *J Virol* **71**: 8087–95. http://www.ncbi.nlm.nih.gov/pubmed/9343157%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC192263.

Malone CD, Hannon GJ. 2009. Small RNAs as Guardians of the Genome. *Cell* **136**: 656–668.

Manno G La, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, et al. 2018. RNA velocity of single cells. *Nature* **560**: 494–498.

Maragkakis M, Alexiou P, Nakaya T, Mourelatos Z. 2016. CLIPSeqTools — a novel bioinformatics CLIP-seq analysis suite. *RNA* **22**: 1–9.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12. http://journal.embnet.org/index.php/embnetjournal/article/view/200.

McKinney W. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (eds. S. Van der Walt and J. Millman), pp. 51–56.

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends Genet* **23**: 183–191.

Miyoshi N, Stel JM, Shioda K, Qu N, Odahima J, Mitsunaga S, Zhang X, Nagano M, Hochedlinger K, Isselbacher KJ, et al. 2016. Erasure of DNA methylation, genomic imprints, and epimutations in a primordial germ-cell model derived from mouse pluripotent stem cells. *Proc Natl Acad Sci U S A* **113**: 9545–9550.

Moore LD, Le T, Fan G. 2013. DNA methylation and its basic function. *Neuropsychopharmacology* **38**: 23–38.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**:

621–628.

Müller S, Rycak L, Winter P, Kahl G, Koch I, Rotter B. 2013. omiRas: A Web server for differential expression analysis of miRNAs derived from small RNA-Seq data. *Bioinformatics* **29**: 2651–2652.

Nagosa S, Leesch F, Putin D, Bhattacharya S, Altshuler A, Serror L, Amitai-Lange A, Nasser W, Aberdam E, Rouleau M, et al. 2017. microRNA-184 Induces a Commitment Switch to Epidermal Differentiation. *Stem Cell Reports* **9**: 1991–2004. https://doi.org/10.1016/j.stemcr.2017.10.030.

Nakamura T, Yabuta Y, Okamoto I, Aramaki S, Yokobayashi S, Kurimoto K, Sekiguchi K, Nakagawa M, Yamamoto T, Saitou M. 2015. SC3-seq : a method for highly parallel and quantitative measurement of single-cell gene expression. *Nucleic Acids Res* **43**: 1–17.

Nakato R, Itoh T, Shirahige K. 2013. DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data. *Genes to Cells* **18**: 589–601.

Navarro FC, Hoops J, Bellfy L, Cerveira E, Zhu Q, Zhang C, Lee C, Gerstein MB. 2019. TeXP: Deconvolving the effects of pervasive and autonomous transcription of transposable elements. *PLOS Comput Biol* **15**: 1–19.

Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, Lu Z, Stitzer MC, Crisp PA, Hirsch CN, Zhang X, et al. 2019. Monitoring the interplay between transposable element families and DNA methylation in maize. *PLoS Genet* 1–25.

Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M. 2013. MIPS PlantsDB : a database framework for comparative plant genome research. *Nucleic Acids Res* **41**: 1144–1151.

O'Neill K, Liao WW, Patel A, Hammell M. 2018. TEsmall identifies small RNAs associated with targeted inhibitor resistance in melanoma. *Front Genet*.

Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC. 2007. The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. *Cell* **130**: 89–100.

Pace II JK, Feschotte C. 2007. The evolutionary history of human DNA transposons : Evidence for intense activity in the primate lineage. *Genome Res* 422–432.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417–419.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**: 2825–2830.

Percharde M, Lin C-J, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M. 2018. A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**: 391–405.

Pinnix CC, Herlyn M. 2007. The many faces of Notch signaling in skin-derived cells. *Pigment Cell Res* **20**: 458–465.

Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, Theunissen TW, Jaenisch R, Trono D. 2019. Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell* **24**: 724-735.e5.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rahman RU, Gautam A, Bethune J, Sattar A, Fiosins M, Magruder DS, Capece V, Shomroni O, Bonn S. 2018. Oasis 2: Improved online analysis of small RNA-seq data. *BMC Bioinformatics* **19**: 1–10.

Raviram R, Rocha PP, Luo VM, Swanzey E, Miraldi ER, Chuong EB, Feschotte C, Bonneau R, Skok JA. 2018. Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biol* **19**: 1–19.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015.

Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**: e47.

Robinson JL, Lee EB, Xie SX, Rennert L, Suh E, Bredenberg C, Caswell C, Van Deerlin VM, Yan N, Yousef A, et al. 2018. Neurodegenerative disease concomitant proteinopathies are prevalent, age-related and APOE4-associated. *Brain* **141**: 2181–2193.

Rodriguez-Terrones D, Gaume X, Ishiuchi T, Weiss A, Kopp A, Kruse K, Penning A, Vaquerizas JM, Brino L, Torres-Padilla M-E. 2018. A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat Genet* **50**.

Rowe HM, Kapopoulou A, Corsinotti A, Fasching L, Macfarlan TS, Tarabay Y, Viville S, Jakobsson J, Pfaff SL, Trono D. 2013. TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Res* **23**: 452–461.

Rueda A, Barturen G, Lebrón R, Gómez-Martín C, Alganza Á, Oliver JL, Hackenberg M. 2015. SRNAtoolbox: An integrated collection of small RNA research tools. *Nucleic Acids Res* **43**: W467–W473.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves Â, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348.

Schorn AJ, Gutbrod MJ, LeBlanc C, Martienssen R. 2017. LTR-Retrotransposon Control by tRNA-Derived Small RNAs. *Cell* **170**: 61–71. http://dx.doi.org/10.1016/j.cell.2017.06.013.

Schorn AJ, Martienssen R. 2018. Tie-Break: Host and Retrotransposons Play tRNA. *Trends Cell Biol* **28**: 793–806. https://doi.org/10.1016/j.tcb.2018.05.006.

Seczynska M, Bloor S, Cuesta SM, Lehner PJ. 2022. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature* **601**: 440–445.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Haeseler A Von, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**: 461–468.

Sexton CE, Han M V. 2019. Paired-end mappability of transposable elements in the human genome. *Mob DNA* **10**: 1–11.

Sheen F, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD. 2000. Reading between the LINEs : Human Genomic Variation Induced by LINE-1 Retrotransposition. *Genome Res* **10**: 1496–1508.

Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, Brennan PM, Baillie JK, Collino A, Ghisletti S, et al. 2013. Endogenous Retrotransposition Activates Oncogenic Pathways in Hepatocellular Carcinoma. *Cell* **153**: 101–111.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**: 272–285.

Soifer HS, Zaragoza A, Peyvan M, Behlke MA, Rossi JJ. 2005. A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. *Nucleic Acids Res* **33**: 846–856.

Sovic I, Sikic M, Wilm A, Fenlon SN, Chen S, Nagarajan N. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* **7**.

Stuart T, Eichten SR, Cahn J, Karpievitch Y V, Borevitz JO, Lister R. 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife* **5**: 1–27.

Suk TR, Rousseaux MWC. 2020. The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. *Mol Neurodegener* **15**: 1–16.

Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, Li W. 2014. MOABS : model based analysis of bisulfite sequencing data. *Genome Biol* **15**: 1–12.

Sun G, Chung D, Liang K, Keles S. 2013. Statistical Analysis of ChIP-seq Data with MOSAiCS. In *Deep Sequencing Data Analysis* (ed. N. Shomron), pp. 193–212, Springer Science+Business Media, New York, NY.

Sun X, Wang X, Tang Z, Grivainis M, Kahler D, Yun C, Mita P, Fenyö D, Boeke JD. 2018. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc Natl Acad Sci* **115**.

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1–15.

Svoboda P, Stein P, Anger M, Bernstein E, Hannon GJ, Schultz RM. 2004. RNAi and expression of retrotransposons MuERV-L and IAP in preimplantation mouse embryos. *Dev Biol* **269**: 276–285.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**: 534–538.

Tam OH, Ostrow LW, Gale Hammell M. 2019a. Diseases of the nERVous system: Retrotransposon activity in neurodegenerative disease. *Mob DNA* **10**: 1–14.

Tam OH, Rozhkov N V., Shaw R, Kim D, Hubbard I, Fennessey S, Propp N, Phatnani H, Kwan J, Sareen D, et al. 2019b. Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Rep* **29**: 1164-1177.e5.

Tanay A, Regev A. 2017. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**: 331–338.

Tokuyama M, Kong Y, Song E, Jayewickreme T, Kang I, Iwasaki A. 2018. ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc Natl Acad Sci* **115**.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJ Van, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 516–520.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing : computational challenges and solutions. *Nat Rev Genet* **13**.

Trizzino M, Park YS, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res* **27**: 1623–1633.

Valdebenito-Maturana B, Riadi G. 2018. TEcandidates: prediction of genomic origin of expressed transposable elements using RNA-seq data.

Bioinformatics **34**: 3915–3916.

Van Allen EM, Wagle N, Sucker A, Treacy DJ, Johannessen CM, Goetz EM, Place CS, Taylor-Weiner A, Whittaker S, Kryukov G V., et al. 2014. The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discov* **4**: 94–109.

Vanden Broeck L, Callaerts P, Dermaut B. 2014. TDP-43-mediated neurodegeneration: Towards a loss-of-function hypothesis? *Trends Mol Med* **20**: 66–71. http://dx.doi.org/10.1016/j.molmed.2013.11.003.

Vangeel EB, Izzi B, Hompes T, Vansteelandt K, Lambrechts D, Freson K, Claes S. 2015. DNA methylation in imprinted genes IGF2 and GNASXL is associated with prenatal maternal stress. *Genes, Brain Behav* **14**: 573–582.

Venuto D, Bourque G. 2018. Identifying co-opted transposable elements using comparative epigenomics. *Dev Growth Differ* **60**: 53–62.

Villanueva J, Vultur A, Herlyn M. 2011. Resistance to BRAF Inhibitors : Unraveling Mechanisms and Future Treatment Options. *Cancer Res* **71**: 7137–7141.

Villanueva J, Vultur A, Lee JT, Somasundaram R, Cipolla AK, Wubbenhorst B, Xu X, Phyllis A, Kee D, Santiago-walker AE, et al. 2010. Acquired resistance to BRAF inhibitors mediated by a RAF kinase switch in melanoma can be overcome by co-targeting MEK and IGF-1R/PI3K.

Cancer Cell **18**: 683–695.

Vitsios DM, Enright AJ. 2015. Chimira: Analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics* **31**: 3365–3367.

Wang J, Huda A, Lunyak V V, Jordan IK. 2010. A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics* **26**: 2501–2508.

Wang K, Liang C, Liu J, Xiao H, Huang S, Xu J, Li F. 2014. Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics* **15**: 1–8.

Wang R, Hsu H, Blattler A, Wang Y, Lan X, Wang Y, Hsu P-Y, Leu Y-W, Huang TH, Farnham PJ, et al. 2013. LOcating Non-Unique matched Tags (LONUT) to Improve the Detection of the Enriched Regions for ChIP-seq Data. *PLoS One* **8**: 1–10.

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* **104**: 18613–18618.

Warnes GR, Bolker B, Lumley T. gplots: Various R programming tools for plotting data. R package version 2.6.0.

Waters S, Swanson MEV, Dieriks B V., Zhang YB, Grimsey NL, Murray HC, Turner C, Waldvogel HJ, Faull RLM, An J, et al. 2021. Blood-spinal cord barrier leakage is independent of motor neuron pathology in ALS. *Acta Neuropathol Commun* **9**: 1–17. https://doi.org/10.1186/s40478-021-01244-0.

Wicker T, Matthews DE, Keller B. 2002. TREP : a database for Triticeae repetitive elements. *Trends Plant Sci* **7**: 561–562.

Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York http://ggplot2.org.

Wilson R, Doudna JA. 2013. Molecular mechanisms of RNA interference. *Annu Rev Biophys* **42**: 217–239. http://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC5895182&blobtype=pdf.

Wong TN, Miller CA, Jotte MRM, Bagegni N, Baty JD, Schmidt AP, Cashen AF, Duncavage EJ, Helton NM, Fiala M, et al. 2018. Cellular stressors contribute to the expansion of hematopoetic clones of varying leukemic potential. *Nat Commun* **9**: 1–10.

Workshop TBW. 2019. Muangthai “The Elbow King”"Elbow Zombie Highlight". *YouTube*.

Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In *Statistical Genomics: Methods and Protocols* (eds. E. Mathé and S. Davis), pp. 283–334, Springer New York, New York, NY.

Xi Y, Li W. 2009. BSMAP : whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**: 1–9.

Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet* **45**: 836–841.

Yang A, Tang JYS, Troup M, Ho JWK. 2019a. Scavenger: A pipeline for recovery of unaligned reads utilising similarity with aligned reads. *F1000 Res* 1–20.

Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. 2019b. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res* **47**: 1–16.

Yi F, Jia Z, Xiao Y, Ma W, Wang J. 2018a. SPTEdb : a database for transposable elements in salicaceous plants. *Database* 1–8.

Yi F, Ling J, Xiao Y, Zhang H, Ouyang F, Wang J. 2018b. ConTEdb : a comprehensive database of transposable elements in conifers. *Database* 1–7.

Zeng X, Li B, Welch R, Rojo C, Zheng Y, Dewey CN, Keles S. 2015. Perm-seq : Mapping Protein-DNA Interactions in Segmental Duplication and Highly Repetitive Regions of Genomes with Prior-Enhanced Read Mapping. *PLoS Comput Biol* **11**: 1–23.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**.

Zhang T, Cooper S, Brockdorff N. 2015. The interplay of histone modifications – writers that read. *EMBO Rep* **16**: 1467–1481.

Zhang Y, Wang X, Kang L. 2011. A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics* **27**: 771–776.

Zhang Z, Xing Y. 2017. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res* **45**: 9260–9271.

Zheng Y, Ay F, Keles S. 2019. Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. *Elife* 1–29.

Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W. 2017. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**: 631–643.

Zlotorynski E. 2014. RNA interference: MicroRNAs suppress transposons. *Nat Rev Mol Cell Biol* **15**: 298–299. http://dx.doi.org/10.1038/nrm3788.

Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. 2019. A primer on deep learning in genomics. *Nat Genet* **51**: 12–18. http://dx.doi.org/10.1038/s41588-018-0295-5.