NORTHWESTERN UNIVERSITY

Molecular and Computational Studies of TDP-43 and FUS Proteinopathies

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Neuroscience

By Warren McGee

EVANSTON, ILLINOIS

June 2019

ABSTRACT

Frontotemporal Dementia (FTD) and Amyotrophic Lateral Sclerosis (ALS) are two devastating neurodegenerative diseases that affect 100,000s of people globally. They have a severe adverse impact on society, yet there are currently no early diagnostic tools or disease-modifying therapies available. Despite their clinical heterogeneity, evidence points to these diseases being on a spectrum, with shared molecular characteristics.

Two proteins known to be associated with the disease spectrum are TDP-43 and FUS, both multifunctional DNA- and RNA-binding proteins. These two proteins share common structural features, have a core of common target genes, and have similar functions throughout the lifecycle of RNA, regulating transcription, splicing, localization, translation and stability. However, they also have distinct characteristics and differences as well. For both proteins, the current paradigm says that a combination of factors leads to nuclear clearance and aggregation into cytosolic inclusion bodies. An unresolved debate in the field is whether the disease occurs through loss-of-function or gain-of-toxicity mechanisms.

This work was motivated to better understand the endogenous roles these proteins play in regulating the nervous system. In particular, much recent work has provided evidence that both proteins are involved with miRNA biogenesis and mitochondrial function, both of which have been implicated in the pathogenesis of the ALS-FTD spectrum. Because of the complexity of the number of potential RNA targets, a genome-wide approach is essential for understanding the roles of these proteins. Thus, we sought to explore both of these functions systematically using a combination of molecular biology and bioinformatics.

We first systematically examined which miRNAs are regulated by TDP-43 and FUS in neuronal model systems. In this work, we designed a novel pipeline to both predict which

miRNA-mRNA interactions are occurring in our model system, but also which pathways might be dysregulated. We identified FUS-regulated miRNAs that had established and predicted roles in synaptic regulation. Intriguingly, we identified several cancer-associated miRNAs regulated by TDP-43, with several TDP-43-regulated miRNAs predicted to have novel roles in lung cancer pathogenesis and prognosis. In particular, our pipeline identified one miRNA, miR-423-3p, with a predicted role in regulating cell migration. Follow-up experiments validated this prediction, demonstrating the power of this network approach.

We next sought to determine which mitochondrial-associated genes may be regulated by FUS. In the course of initial work in this area, we realized that RNA-Sequencing data is compositional rather than count data. This means that it carries only relative information, not information about absolute copy number changes. Standard normalization methods can lead to distorted results if there is significantly more RNA in one condition versus another. We thus designed a new normalization approach, called compositional normalization (implemented in an extension of the popular sleuth tool, called **sleuth-ALR**), to deal with this problem, and we also designed a new simulation protocol, **absSimSeq**, to benchmark performance in a more accurate way. Compositional normalization performed similarly to standard normalization when analyzing data that did not have large number of changes; however, compositional normalization had much improved performance when analyzing data that did have a large number of changes. Applying sleuth-ALR to FUS RNA-Sequencing datasets led to a dramatic re-interpretation of the global expression patterns, as well as which set of transcripts and pathways were dysregulated.

Finally, we studied the role of FUS in mitochondrial function in an HEK CRISPR/Cas9 KO model. We did not observe any changes in proliferation, bioenergetics, or mitochondrial membrane potential measured at the cell-level. Surprisingly, we observed an increase in the

membrane potential of purified mitochondria, and we also present preliminary evidence that several mitochondrial-associated transcripts are up-regulated after FUS KO, suggesting the intriguing possibility that FUS is an inhibitor of mitochondrial function.

Acknowledgments

This work could not have been possible without the help of so many people. First of all, my mentor, Dr. Jane Wu, deserves a large chunk of the praise. From countless hours providing guidance and support, setting a high standard but helping me to meet it, and even being there for me in the bad times—she always drove me to my full potential and showed a great dedication to the success and wellbeing of all her trainees. I will treasure my time under her tutelage.

To Dr. Fushimi, Xiaoping, Ryan, Mickie, and all of the other Wu Lab members: thank you for the technical support, for the laughs, for the exchange of food, culture and life, and for just making the time in the lab such a fun experience. To Dr. Li Zhu, Dr. Jianghong Liu, Dr. Jianwen Deng, Dr. Peng Wang, Dr. Ruirui Kong, and all of the other members of the IBP group: thank you for welcoming me into your lab for those two months, for imparting to me your technical expertise, for giving me your hospitality and friendship, and for sharing with me all the wonders of the Chinese culture. To my committee—Dr. Kessler, Dr. Chandel, Dr. Kording, and Dr. Braun: thank you for all of your mentorship and support, especially for when you had to give me important and constructive feedback (on my work, my career, and life), and it was tough to swallow. To my fellow peers who became my teachers, especially Daniel Fisher, Sam Weinberg, and Manan Mehta: without your support, I would have been lost at the bench. To my collaborators, Harold Pimentel and Lior Pachter: thank you for taking a chance on a lone wolf bioinformatician and helping bring my computational work to the finish line.

Finally, to my family, my siblings, friends and my dearest Kathryn: thank you for all of your love and support over this long, arduous journey. When I doubted, when I wasn't taking care of myself, and when I lost sight of the goal, you were right there to help me stay balanced, help me persevere, and help remind me of the joy, love, and vocation underneath it all.

To Him Who made all of this possible

A.M.D.G.

TABLE OF CONTENTS

List of Figures and Tables

MAIN FIGURES AND TABLE

APPENDIX FIGURES

PREFACE

Parts of this dissertation did appear in the following publications:

Chen, X., Fan, Z., **McGee, W.**, Chen, M., Kong, R., Wen, P., Xiao, T., Chen, X., Liu, J., Zhu, L., et al. (2018). TDP-43 regulates cancer-associated microRNAs. Protein & Cell 9, 848–866.

Parts of this dissertation will appear in:

McGee, W.A., Pimentel, H., Pachter, L., and Wu, J.Y. (2019). Compositional Data Analysis is necessary for simulating and analyzing RNA-Seq data.

CHAPTER 1: INTRODUCTION

1.1 The Neurodegenerative Diseases FTD and ALS

1.1.1 The Clinical and Social Impact of FTD and ALS

Frontotemporal Dementia (FTD) is the second most common cause of early-onset dementia (<65 y/o), with an estimated incidence of 2-5/100,000 patients and a prevalence of 15-22/100,000 patients (Onyike and Diehl-Schmid, 2013; Rabinovici and Miller, 2010; Seltman and Matthews, 2012). This means that FTD affects roughly 20,000-30,000 adults in the US (Knopman and Roberts, 2011), with ~60% of patients developing the disease between the ages of 45-65 and ~13% developing it before age 50 (Onyike and Diehl-Schmid, 2013). Amyotrophic Lateral Sclerosis (ALS) is the most common motor neuron disease, with a global incidence ranging from 0.5-3/100,000 people (Al-Chalabi and Hardiman, 2013; Beghi et al., 2006; Hardiman et al., 2017). This translates to the disease affecting roughly 15,000-20,000 people in the US, with approximately 5,000 new US cases diagnosed every year (Mehta, 2018).

Like other neurodegenerative diseases, FTD and ALS both represent a substantial economic and social burden (Galvin et al., 2017; Larkindale et al., 2014; Shrestha and Heisler, 2011). Specifically, FTD costs on average $120,000 per year for affected families (including both direct and indirect costs), which is double the typical costs for a family affected by Alzheimer's dementia (Galvin et al., 2017). ALS costs approximately $64,000 per year for affected families, and has an estimated national burden of roughly $1 billion per year (Larkindale et al., 2014). Both diseases primarily affect patients between the third and sixth decades of life, meaning that these patients would otherwise be in the workforce. This represents a special burden compared to Alzheimer's dementia, which typically affects patients in the later decades.

Importantly, there are no therapies available for FTD, and only one approved therapy for ALS (riluzole). There are no disease-modifying therapies nor any diagnostic tests for early detection for either disease.

1.1.2 Clinical Features of FTD and ALS

1.1.2.a FTD

FTD defines a heterogeneous group of clinical syndromes characterized by progressive and selective loss of the executive, behavioral, and language cognitive domains (Erkkinen et al., 2018; Kaivorinne, 2012; Seltman and Matthews, 2012; Sieben et al., 2012; Young et al., 2017). The prognosis varies widely, with a median survival of 6-12 years from symptom onset and 3-4 years from diagnosis (Kansal et al., 2016; Ng et al., 2015). Usually, though, the clinical course is more aggressive than what is seen in Alzheimer's dementia (Rascovsky et al., 2005).

There are two broad clinical categories based on the predominant clinical features and pattern of atrophy: the behavioral variant, called behavioral variant FTD (bvFTD), and the language variant, called Primary Progressive Aphasia (PPA). PPA can be subdivided further based on the divergent localization and precise language problems that develop: semantic variant PPA (i.e. Semantic Dementia, SD); non-fluent variant PPA (i.e. Progressive Non-Fluent Aphasia, PNFA), and logopenic variant PPA (Erkkinen et al., 2018; Mackenzie et al., 2008; Young et al., 2017). There are additional uncommon variants that are also included under the umbrella, which are reviewed elsewhere (Erkkinen et al., 2018; Kaivorinne, 2012).

There appears to be a strong genetic component, with ~25-50% of patients having a family history of dementia, and ~10% showing an autosomal dominant pattern of inheritance (Rohrer et al., 2009). Three major genes have been identified to have mutations that cause the

disease (in order of discovery: MAPT/tau, GRN, and C9orf72); a dozen or so more genes have been linked with the disease in rare cases (reviewed in (Pottier et al., 2016)).

1.1.2.b ALS

In contrast, ALS is clinically characterized by a progressive loss of motor neurons, with a median survival of 3-5 years from diagnosis (Al-Chalabi and Hardiman, 2013; Bäumer et al., 2014). The clinical manifestations are varied, with patients initially presenting with upper motor neuron dysfunction, lower motor neuron dysfunction, or with a bulbar onset (Hardiman et al., 2017). About 10% of cases have a family history of ALS, with the rest being considered sporadic (Chiò et al., 2008; Renton et al., 2014; Rowland and Shneider, 2001). Since SOD1 was first gene identified to cause ALS (Rosen et al., 1993), more than twenty genes have been identified as causative for ALS (Al-Chalabi et al., 2012; Zou et al., 2017). The most commonly mutated gene is C9orf72, accounting for 10-15% of all ALS cases, familial and sporadic (Al-Chalabi et al., 2017).

1.1.2.c FTD-ALS spectrum

Importantly, there are subset of patients with ALS or FTD that go on to develop the other syndrome; specifically, with approximately ~15% of FTD patients developing ALS, and approximately 30% of ALS patients developing FTD (Lomen-Hoerth, 2011; Rascovsky et al., 2011). This, along with other genetic and pathological evidence, indicates that the two diseases exist on a spectrum (Gao et al., 2017; Geser et al., 2010; 2009). Many pathways have been implicated as dysregulated in this spectrum, including proteostasis, RNA metabolism, RNA and protein transport (including nucleocytoplasmic transport), and inflammation. Two processes of particular relevance to this work are the roles of microRNA dysfunction and mitochondrial dysfunction in these diseases.

1.1.3 Dysregulation of miRNA biogenesis in ALS and FTD

MicroRNAs (miRNAs) are short noncoding RNAs (15-34 nt in length, average 22 nt) with a complex biogenesis and are involved in regulating almost every biological process (Finnegan and Pasquinelli, 2013; Ryan et al., 2015; Wilczynska and Bushell, 2015). Briefly, the canonical pathway consists of transcription of a primary miRNA transcript (pri-miRNA) followed by processing into a hairpin pre-miRNA (~60-70 nt) by the microprocessor complex, composed of Drosha and DGCR8. This pre-miRNA is then exported into the cytoplasm by exportin-5. This is processed further to a duplex of mature miRNAs by another complex composed of Dicer and TRBP. Finally, chaperone proteins like HSP90 and other proteins help load one of the two strands into an Argonaute protein (Ago1-4 in mammals) to form an RNA silencing complex (RISC) (Czech and Hannon, 2011).

Recent studies have implicated miRNAs in neurodegenerative disorders, specifically ALS and FTD (reviewed in (Eitan and Hornstein, 2016; Gascon and Gao, 2012; 2014; Rinchetti et al., 2017)). Conditional knockout of Dicer in spinal motor neurons, thus blocking the production of miRNAs in these neurons, leads to dysfunction and loss of these neurons and recapitulated many pathological features of ALS (Haramati et al., 2010). On the other hand, enhancing Dicer function using the small molecule enoxacin partially alleviated neuromuscular dysfunction in two ALS mouse models (Emde et al., 2015). In addition, Drosha and DGCR8 were found in inclusions co-localizing with the dipeptide repeats produced by C9orf72 mutations, in both FTD and ALS patients (Porta et al., 2015). Finally, using microarrays, multiple miRNAs have been found to be dysregulated in FTD and ALS patients (Campos-Melo et al., 2013; Figueroa-Romero et al., 2016; Kocerha et al., 2011).

1.1.4 Mitochondrial dysfunction in ALS and FTD

Mitochondria are critical organelles, both for energy homeostasis (through oxidative phosphorylation) and for integrating multiple signals in the cell (Bohovych and Khalimonchuk, 2016; Chandel, 2014). These functions are especially critical in neurons, where great distances separate distal synaptic terminals from the soma and maintaining synaptic transmission has exceptionally high energy demands (Sheng, 2017). Of particular importance is so-called "retrograde signaling", where mitochondria send signals back to the nucleus to modulate transcription (Hunt and Bateman, 2018). Besides the endoplasmic reticulum, mitochondria also serve as an additional site for calcium uptake and buffering; this is important in neurons, which use calcium signaling extensively during synaptic transmission (Pivovarova and Andrews, 2010). Interestingly, there are different populations of mitochondria in neurons (somatic mitochondria versus synaptic mitochondria) which have different properties related to the rate of the complexes that participate in oxidative phosphorylation (Davey and Clark, 1996; Davey et al., 1998) and the capacity to buffer calcium (Brown et al., 2006; Stauch et al., 2014).

Mitochondrial dysfunction has been known to be an aspect of ALS for a long time (Carrí et al., 2015; Cozzolino and Carrí, 2012; Jiang et al., 2015). Histological examination of ALS patient neurons showed swollen and vacuolated mitochondria (Cozzolino and Carrí, 2012). The first gene identified as mutated in ALS, SOD1, is a mitochondrial gene; much of the early work after its discovery was focused on the role of oxidative stress in ALS (reviewed in (Tan et al., 2014)). Since then, multiple studies have observed defects in oxidative phosphorylation in cells taken from ALS patients (reviewed in (Cozzolino and Carrí, 2012)), including fibroblasts taken from patients with the C9orf72 mutation (Onesto et al., 2016). In addition, there is evidence to

link severity of disease in ALS with body weight, lipid profiles, and diabetes status (Jawaid et al., 2018), indicating that metabolism influences the disease.

Though not as well studied, mitochondrial dysfunction has also been implicated in FTD. A novel mutation discovered in rare patients with ALS and FTD was found in CHCHD10, another mitochondrial gene (Bannwarth et al., 2014). Additional rare mutations in p97/VCP and p62/SQSTM1 have been found in patients with FTD, and both have also been implicated in mitochondrial dysfunction (reviewed in (Solomon et al., 2019)). Finally, a recent case series of FTLD post-mortem brain samples used electron microscopy to reveal significant mitochondrial damage, especially a marked loss or disruption of cristae (Deng et al., 2015).

1.2 TDP-43 and FUS proteinopathies

1.2.1 Discovery of the proteinopathies

TAR DNA-Binding Protein, 43 kDa (TDP-43), was first identified in the context of repressing the transcription of the HIV-1 trans-activation response element (Ou et al., 1995). The Fused in Sarcoma / Translocated in Liposarcoma gene (FUS/TLS, hereafter FUS) was first discovered in the context of cancer, forming a fusion gene with CHOP to cause malignant myxoid liposarcoma (Crozat et al., 1993; Rabbitts et al., 1993). Both are multifunctional DNA and RNA binding proteins, participating in processes that span all steps of RNA processing from transcription, splicing, localization, and translation (see **Section 1.3** below for a review of their structure and function).

In 2006, TDP-43 was found to be the major protein component of ubiquitin positive inclusion bodies in patients with both Frontotemporal Lobar Degeneration (FTLD) and Amyotrophic Lateral Sclerosis (ALS) (Arai et al., 2006; Neumann et al., 2006). Later, in 2009, two groups discovered mutations in FUS that cause ALS (Kwiatkowski et al., 2009; Vance et al.,

2009), and another group showed FUS to be the major protein component of ubiquitin positive inclusion of another subset of patients with FTLD (Neumann et al., 2009). Since then it has been realized that these cases are part of a clinical spectrum and have been reclassified as "TDP-43 proteinopathy" and "FUS proteinopathy" because of their characteristic feature of FUS-positive inclusion bodies (Geser et al., 2010; Mackenzie and Neumann, 2017a).

1.2.2 Pathological Characteristics of TDP-43 and FUS proteinopathies

1.2.2.a FTLD

Frontotemporal Lobar Degeneration (FTLD) defines the underlying pathology associated with FTD (Mackenzie and Neumann, 2016; Mackenzie et al., 2009; Neumann and Mackenzie, 2019). The relationships between the two are complex, where one clinical category can be associated with multiple pathologies, and one type of pathology can be associated with multiple clinical syndromes (Erkkinen et al., 2018; Mackenzie and Neumann, 2016; Mackenzie et al., 2009; Neumann and Mackenzie, 2019). The most common pathology is FTLD pathology due to TDP-43 (FTLD-TDP), accounting for ~45-50% of cases; the next most common is pathology due to the AD-associated protein tau (FTLD-Tau), accounting for ~40-45% of cases; third is pathology due to FUS (FTLD-FET), accounting for ~10% of cases (Kaivorinne, 2012).

The normal histological pattern of TDP-43 immunostaining is a diffuse distribution predominantly in the nucleus. In contrast, the hallmark pathological pattern found in FTLD-TDP cases is a nuclear clearance accompanied by dense aggregations that are immunoreactive to TDP-43 and ubiquitin (Neumann and Mackenzie, 2019). These lesions vary widely by morphology and location, but frequently identified lesions include neuronal cytoplasmic inclusions, intranuclear inclusions, and dystrophic neurites. Despite the variability,

neuropathologists have found common patterns that led to a consensus definition of four pathological subtypes of FTLD-TDP (Mackenzie et al., 2011; Mackenzie and Neumann, 2017b).

More recent work, especially with bvFTD patients, have identified specific regions and specific subpopulations of neurons that are especially vulnerable to developing TDP-43 pathology (Nana et al., 2018; Seeley, 2008). In bvFTD, the regions most affected include the anterior cingulate and frontoinsular cortices, but other frontotemporal cortices and subcortical structures frequently have pathological lesions (Nana et al., 2018; Neumann and Mackenzie, 2019). Intriguingly, a recent case series with identified neurons that had nuclear clearance of TDP-43 without any discernible aggregations of TDP-43, yet these neurons had similar atrophy compared to neurons that had TDP-43 aggregations (Nana et al., 2018).

Similar to TDP-43, the normal histological pattern of FUS immunostaining is also a predominantly nuclear and diffuse staining (Mackenzie and Neumann, 2017a). In FTLD-FET, the characteristic lesions are also neuronal cytoplasmic inclusions that are strongly immunoreactive for FUS and ubiquitin, frequently (but not always) accompanied by nuclear clearance of FUS (Mackenzie and Neumann, 2017a). These cases are called FTLD-FET because FUS inclusions frequently also contain EWSR1 and TAF15, which with FUS form the FET family or proteins; these lesions also frequently include Transportin1, the carrier protein that facilitates transport of the FET proteins out of the nucleus (Mackenzie and Neumann, 2017a). In contrast to FTLD-TDP, the caudate nucleus is characteristically affected in FTLD-FET (Josephs et al., 2010).

1.2.2.b ALS

In ALS, the underlying pathology in a large majority of cases (95-98%) contains TDP-43 (ALS-TDP) (Mackenzie et al., 2007), with the rest of the cases containing pathology either for

SOD1 (ALS-SOD1, in patients with SOD1 mutations) and for FUS (ALS-FUS, in patients with FUS mutations) (Mackenzie et al., 2010). Interestingly, though TDP-43 and FUS mutations have been identified in ALS-TDP and ALS-FUS cases, respectively, so far TDP-43 and FUS mutations have only been rarely detected in FTLD-FET cases (Kaivorinne, 2012; Pratt et al., 2012).

Similar to FTLD-TDP, the characteristic lesion in ALS-TDP patients is a clearance of nuclear TDP-43 accompanied by nuclear cytoplasmic inclusions immunoreactive to TDP-43 and ubiquitin (Saberi et al., 2015). These lesions are found in lower and upper motor neurons, along with the frontal and temporal cortices, the hippocampus, and striatum.

In addition, motor neurons in ALS-TDP frequently have an additional lesion called a "Bunina body", small oval-to-round eosinophilic lesions found intracellularly in both lower and upper motor neurons (Saberi et al., 2015). They frequently are positive for Transferrin and Cystatin C, but are negative for other proteins frequently associated with neurodegeneration (Saberi et al., 2015). The biological significance is unknown. Interestingly, some subpopulations of motor neurons only rarely have Bunina bodies: Betz cells (the large projection neurons from the motor cortex to the spinal cord), the oculomotor nuclei, and the Onuf nuclei (found in the sacral spinal cord) (Saberi et al., 2015). Importantly, whereas the oculomotor nuclei and the Onuf nuclei are typically spared in ALS (Brockington et al., 2012; Fogarty, 2018), Betz cells are frequently atrophied and lost in ALS (Braak et al., 2017; Fogarty, 2018). A recent case series of ALS-TDP patients observed that Betz cells frequently have TDP-43 nuclear clearance without accompanying cytoplasmic inclusions (Braak et al., 2017), suggesting that Bunina bodies are related to the cytoplasmic aggregations.

Similar to ALS-TDP, patients with ALS-FUS also have a characteristic nuclear clearance of FUS with accompanying neuronal cytoplasmic inclusions (Mackenzie and Neumann, 2017a). Previous work found that the degree of nuclear clearance of FUS was associated with the severity of the disease, as characterized by age of onset and rate of progression (Dormann et al., 2010; Mackenzie and Neumann, 2017a). In contrast to ALS-TDP, there are no Bunina bodies in the neurons of ALS-FUS patients (Mackenzie and Neumann, 2017a). In contrast to FTLD-FET, FUS inclusions in ALS-FUS patients do not co-localize with EWSR1, TAF15, or Transportin1 (Mackenzie and Neumann, 2017a).

These differences suggest that, whereas nuclear clearance is a common theme across FUS and TDP-43 proteinopathies, there are likely different mechanisms in specific groups of patients with these diseases. A detailed understanding of the pathogenesis of these diseases is needed to better understand how to diagnose and treat these diseases. This would be helped by better understanding the structure and function of both TDP-43 and FUS. There are some important similarities between the two, but also some important differences (summarized below in section 1.3).

1.3 The Structure and Function of TDP-43 and FUS

1.3.1 TDP-43

1.3.1.a Structure and function of TDP-43

Interested readers can explore two recent reviews for more in-depth discussion (Guo and Shorter, 2017; Prasad et al., 2019), but here will be a brief review of the structure and function of TDP-43. TDP-43 is a member of the heteronuclear ribonucleoproteins (hnRNP) family, with an N-terminal domain, two RNA Recognition Motifs (RRMs), and a C-terminal “prion-like” domain (Buratti and Baralle, 2001). It has a canonical bipartite nuclear localization signal (NLS)

between the N-terminal domain and the first RRM (aa82-95) (Winton et al., 2008). Though a nuclear export signal (NES) was proposed within the second RRM (aa239-250) (Winton et al., 2008), later work provided strong evidence that this is not a true export signal and that instead TDP-43 likely is exported through passive diffusion (Archbold et al., 2018; Ederle et al., 2018; Pinarbasi et al., 2018).

The N-terminal domain has two conformational states in equilibrium: a highly disordered unfolded state, and a well-folded state. The latter was found by NMR and circular dichroism to be a ubiquitin-like fold, despite low sequence similarity with ubiquitin (Qin et al., 2014). This N-terminal domain can bind to ssDNA (Zhang et al., 2013a), and is required to facilitate homodimerization, an important prerequisite for splicing activity (Jiang et al., 2017; Zhang et al., 2013a). Whereas a truncated form of TDP-43 lacking this domain and the NLS can easily aggregate (Jiang et al., 2017), full-length TDP-43 is less prone to aggregate unless its NLS is mutated (Zhang et al., 2013a). In the context of full-length TDP-43, the extreme N-terminus (residues 1-10) are required for aggregation (Zhang et al., 2013a).

The RRM domains are both required for high-affinity RNA-binding activity, though both can bind RNA independently (Buratti and Baralle, 2001; Lukavsky et al., 2013). Studies have indicated that TDP-43 has a preference for TG/UG-rich motifs (Polymenidou et al., 2011; Tollervey et al., 2011), though UG repeats are neither necessary nor sufficient for TDP-43 binding in cells (Kuo et al., 2014; Polymenidou et al., 2011). An NMR structure containing the two RRMs in tandem with a GU-rich RNA revealed conserved phenylalanines in both domains interacting with three UG repeats, including a central G that stabilizes a conformation between the two domains (Lukavsky et al., 2013). Further work indicated that RRM1 can bind ssDNA by

itself, whereas RRM2 cannot (Furukawa et al., 2016; Kuo et al., 2014); on the other hand, RRM2 was required to provide sequence specificity to the binding of RRM1 (Furukawa et al., 2016).

The majority of TDP-43 mutations identified in ALS are concentrated in the C-terminal domain, suggesting that ALS pathogenesis is at least in part mediated by abnormal aggregation of TDP-43 (Buratti, 2015). As mentioned before, this domain has a high propensity for aggregation as an isolated domain (Jiang et al., 2017). However, this domain is critically important for protein-protein interactions, and for mediating TDP-43's functions in the cell, including splicing and formation of stress granules (Guo and Shorter, 2017; Prasad et al., 2019).

1.3.1.b Models of TDP-43 proteinopathy

A number of cellular and animal models have been developed for TDP-43 proteinopathy across multiple species, from yeast to human (reviewed in (De Giorgio et al., 2019; Guo et al., 2017; Monahan et al., 2018; Romano et al., 2012; Solomon et al., 2019)). Previous work in the Wu lab has produced several models for studying TDP-43, including drosophila models (Li et al., 2010), cultured cortical neurons (Barmada et al., 2010), and stable cell lines expressing wild-type and mutant TDP-43 (Guo et al., 2011; Zhu et al., 2014). TDP-43 is essential for early embryonic development for zebrafish (Schmid et al., 2013) and mice (Kraemer et al., 2010; Sephton et al., 2010), and results in severely restricted lifespan in flies (Feiguin et al., 2009). In contrast, it is dispensable in *C. elegans* (Zhang et al., 2012). Overexpression of TDP-43 is also toxic in multiple species, suggesting that TDP-43 expression is tightly controlled (D'Alton et al., 2014). It is not surprising, then, to note that TDP-43 regulates its own expression by binding to its own 3'UTR (Ayala et al., 2011).

There is evidence to suggest that both loss-of-function and gain-of-toxicity mechanisms are at play in the pathogenesis of ALS and FTD. On the one hand, reduction of TDP-43 leads to

synaptic transmission deficits, motor deficits, and neuron loss (Kraemer et al., 2010; Vanden Broeck et al., 2014). On the other hand, overexpression of TDP-43 or expression of mutant TDP-43 also leads to motor and cognitive deficits (D'Alton et al., 2014). In addition, TDP-43 variants that lead to elevated expression have been directly linked to ALS-FTD in human patients (Gitcho et al., 2009). Also, dysregulation of the TDP-43's autoregulation, as seen with ALS mutant M337V, can lead to motor and cognitive deficits (White et al., 2018). Importantly, neither nuclear clearance of TDP-43 nor cytoplasmic aggregation are required for toxicity (De Giorgio et al., 2019).

Through these studies, TDP-43 has been shown to have multiple roles throughout RNA processing, including splicing (Highley et al., 2014; Polymenidou et al., 2011), suppression of toxic double-stranded RNA and RNAs derived from transposons (Krug et al., 2017; Li et al., 2012; Saldi et al., 2014), transcription (Casafont et al., 2009; Hill et al., 2016), translation (Neelagandan et al., 2019), and RNA stability and localization (Alami et al., 2014; Izumikawa et al., 2017; Tank et al., 2018). The following sections will summarize what is known about TDP-43's role in miRNA biogenesis and mitochondrial function.

1.3.1.c The role of TDP-43 in miRNA biogenesis

Recent work has demonstrated that TDP-43 plays an important role in micro-RNA (miRNA) biogenesis (Eitan and Hornstein, 2016; Gascon and Gao, 2014). When the original Drosha/DGCR8 Microprocessor complex was discovered, TDP-43 was identified as an interacting subunit (Gregory et al., 2004). Two later studies found that TDP-43 can directly regulate Drosha. In human SH-SY5Y neuronal-like cells, TDP-43 was shown to directly regulate the stability of Drosha during retinoic acid treatment, with a global effect on miRNA biogenesis (Di Carlo et al., 2013). In mouse N2a neuronal-like cells, a phosphomimetic mutant of TDP-43

(mimicking the pathologically phosphorylated form of TDP-43 identified in inclusion bodies) also adversely affected the stability of Drosha (Kim et al., 2015). Other work identified that TDP-43 also interacts with the Dicer complex (Kawahara and Mieda-Sato, 2012), and it can directly interact with miRNAs at various stages of processing (Buratti et al., 2010; Kawahara and Mieda-Sato, 2012).

Consistent with the above data, TDP-43 appears to regulate a subset of miRNAs that are relevant to neurodegeneration. One study found that, in flies, TDP-43 regulates a miR-9a, a miRNA critical for neuronal function (Zhu et al., 2012). This finding was replicated in rodent neurons and in patient-derived iPSCs (Zhang et al., 2013b). Further, TDP-43 has been shown to regulate the loading of mature miRNAs into RISC (King et al., 2014). Finally, through microarrays, TDP-43 was found to regulate several miRNAs, and that several TDP-43-associated miRNAs have been found to be dysregulated in disease (Freischmidt et al., 2013; Kawahara and Mieda-Sato, 2012; Kocerha et al., 2011; Rinchetti et al., 2017).

1.3.1.d The role of TDP-43 in mitochondrial function

The first observations linking TDP-43 to mitochondrial function were in 2010. One paper observed a loss in body weight accompanied by loss of body fat and changes in the serum lipid profiles of mice after a conditional TDP-43 knockdown (Chiang et al., 2010). Two other groups saw abnormal aggregations of mitochondria in two independent mouse models over-expressing human TDP-43 (Shan et al., 2010; Xu et al., 2010). The body weight and lipid metabolism observations were replicated in an overexpression mouse model (Stribl et al., 2014), and the mitochondrial aggregation was later replicated in a third independent mouse model expressing a TDP-43 mutant (Xu et al., 2011). In all three mouse models, ultrastructural examination via electron microscopy showed large accumulations of mitochondria (Shan et al., 2010; Xu et al.,

2010; 2011). One study observed a depletion of mitochondria in axon terminals (Shan et al., 2010). These results suggested deficits in mitochondrial transport or fission/fusion dynamics.

Indeed, later work in cultured mouse neurons and in Drosophila models indicated that over-expressed wild-type or mutant TDP-43 disrupts the balance of fission and fusion toward fission (Altanbyek et al., 2016; Wang et al., 2013). A recent study demonstrated that overexpression of TDP-43 in the brain increased fusion protein Mitofusin2 in an age-dependent manner (Davis et al., 2018), and overexpression of Mitofusin2 alone recapitulated the cytoplasmic aggregation observed previously (Huang et al., 2007).

Two of the original groups also observed ultrastructural changes indicating mitochondrial dysfunction, including vacuoles and disrupted cristae (Xu et al., 2010; 2011). These ultrastructural changes were recapitulated in a mouse model expressing a different ALS-associated mutant form of TDP-43 (Stribl et al., 2014). These observations suggested a change in processes related to mitochondrial cristae, especially oxidative phosphorylation. Studies looking at this association have had mixed results. On the one hand, TDP-43 was found to directly localize to mitochondria in post-mortem samples from patients with ALS and FTD, as well as HEK cells, patient fibroblasts, and primary rodent neuronal cultures (Wang et al., 2016), as well as in transgenic mice overexpressing hTDP-43 (Wang et al., 2017). On the other hand, another study using different protocols failed to observe TDP-43 mitochondrial localization in patient fibroblasts with the same mutation nor changes in complex 1 activity (Onesto et al., 2016), and another study failed to detect any differences in oxygen consumption or extracellular acidification (a proxy for glycolytic activity) in HEK cells overexpressing or knocking down TDP-43 (Davis et al., 2018).

Further evidence, though, points to a role for TDP-43 in mitochondrial function. Despite the observation of no change in oxygen consumption, overexpression of wild-type or different mutant TDP-43 disrupted complex 1 assembly via translational inhibition of the ND3 and ND6 mitochondrial-encoded subunits (Wang et al., 2016), and that inhibition of TDP-43 import using a small peptide alleviated these deficits (Wang et al., 2016) and improved motor and cognitive deficits in transgenic mice (Wang et al., 2017). In addition, another group observed that TDP-43 stabilizes RNA intermediates of mitochondrial-encoded transcripts (Izumikawa et al., 2017). Finally, overexpression of TDP-43 was found to disrupt an interaction between mitochondria and endoplasmic reticulum, mediated by ER protein VAPB and mitochondrial protein PTPIP51 (Lau et al., 2018; Stoica et al., 2014).

1.3.2 FUS

1.3.1.a Structure and function of FUS

Although it has long been recognized that FUS is a RNA/DNA binding protein involved in multiple gene regulatory processes, the biological function of this protein in the nervous system remains unclear (Sama et al., 2014). Like TDP-43, it is a member of the heteronuclear ribonucleoproteins (hnRNP) family. Also similar to TDP-43, it has a predominant nuclear localization with shuttling between the nucleus and cytoplasm, and the pathologic form is predominantly in the cytoplasm (Baloh, 2012; Lagier-Tourenne et al., 2010). Interested reads can read two in-depth reviews of FUS for more information (Efimova et al., 2017; Sama et al., 2014), but here will be a brief review of its structure and function.

The structure of FUS has some similarities but also some important differences with TDP-43. It also has a prion-like low complexity domain (LCD), but in the N-terminus. It also has an RRM domain, but instead of a second RRM, it instead has a zinc-finger domain (ZFD)

flanked by two DNA/RNA binding RGG domains. Finally, it contains a C-terminal non-canonical nuclear localization signal. Like TDP-43, one group proposed an NES in its RRM domain (aa289-298) (Kino et al., 2011); however, later work showed that this NES is non-functional and that FUS also likely is exported via passive diffusion (Ederle et al., 2018). Finally, also like TDP-43, FUS can regulate its own expression, though through a different mechanism than TDP-43 (Dini Modigliani et al., 2014; Zhou et al., 2013).

The LCD has been shown to have important roles in the formation of "liquid-liquid phase separation". In this process, FUS, TDP-43 and other proteins interact with RNAs to form membrane-less organelles in both the nucleus and cytoplasm to mediate diverse functions (reviewed in (Prasad et al., 2019; St George-Hyslop et al., 2018; Uversky, 2017)). As described in more detail below, the LCD of FUS is also involved in forming higher-order structures with RNA. Finally, multiple post-translational modifications within the LCD influence the ability of FUS to aggregate and induce phase separation, indicating that this process is tightly controlled (Monahan et al., 2017; Rhoads et al., 2018; Shorter, 2017).

Because of the different domain structure in FUS versus TDP-43, it is not surprising that the mode of RNA and DNA binding is different between the two proteins. Whereas TDP-43 appears to have a sequence preference for TG-rich/UG-rich motifs, there has been considerable debate about a motif for FUS binding (Lagier-Tourenne et al., 2012; Wang et al., 2015b). NMR of FUS's RRM in complex with RNA indicated that the domain adopts a canonical structure despite having a significantly different sequence (Liu et al., 2013). Importantly, it lacks several aromatic residues that typically mediate sequence specificity, and instead has a conserved "KK-loop" that is required for RNA binding via electrostatic interactions (Liu et al., 2013). Later work

revealed that all of the domains except the LCD can mediate DNA and RNA binding (Ozdilek et al., 2017; Schwartz et al., 2013; Wang et al., 2015b).

Further, FUS can bind a wide range of different sequences, both single-strand and double-strand DNA and RNA, though FUS's affinity for double-stranded DNA and RNA is weaker than that for single-strand (Ozdilek et al., 2017; Wang et al., 2015b). Instead, the affinity of FUS for RNA appears to be length-dependent. FUS has a binding preference for long introns (Hoell et al., 2011; Ishigaki et al., 2012; Lagier-Tourenne et al., 2012; Rogelj et al., 2012), and it forms fibrils on longer stretches of RNA in a highly cooperative manner; this cooperativity is mediated by the RGG-ZFD-RGG and the LCD (Schwartz et al., 2013; Wang et al., 2015b). It also interacts with general transcription factors and the RNA Polymerase II complex at 1000s of target genes (Sama et al., 2014; Schwartz et al., 2012); it further has a demonstrated role in DNA damage repair, especially damage associated with transcription (Hill et al., 2016; Sama et al., 2014). The emerging picture is of a protein that helps coordinate activities between DNA and RNA from transcription through to localization and translation (Masuda et al., 2016a; Yu and Reed, 2015).

1.3.1.b Models of FUS proteinopathy

Similar to TDP-43, multiple models of FUS proteinopathy have been developed across species, from yeast to humans (De Giorgio et al., 2019; Guo et al., 2017; Lindström and Liu, 2018; Nolan et al., 2016; Solomon et al., 2019). Previous work in the Wu lab has also developed multiple models: in yeast (Chen et al., 2016; 2011; Deng et al., 2015; 2018; Fushimi et al., 2011).

Two FUS KO models were published at the same time in 2000 (Hicks et al., 2000; Kuroda et al., 2000). One KO was done on the C57J/B6 background, and those mice die immediately after birth with unclear etiology(Hicks et al., 2000), though the immune system

appears affected. The second KO was produced on either a 129/SvEvH background, or a mixed outbred CD1 and 129/SvEvH background. The former also had perinatal lethality of unknown cause, but the latter had mice survive to adulthood. However, they had reduced weight, were infertile, and were sensitive to ionizing radiation.

In the Hicks KO model, hippocampal neurons from these mice have abnormal spine development(Fujii et al., 2005). They observed that FUS shuttled to dendrites in an activity-dependent manner, with an increased localization after mGluR5 activation via DHPG. They suggested that FUS localizes RNA to dendritic spines in an activity dependent manner. Another KO model was generated on an outbred line (ICR crossed with B6) that lived to adulthood, and had no motor defects nor tremor, but had behavioral deficits (hyperactivity and reduced anxiety-related behavior), suggesting that the KO does not recapitulate an ALS phenotype, but may recapitulate an FTLD phenotype(Kino et al., 2015). In this outbred model, they observed vacuolation of hippocampal neurons, with the vacuoles staining positive for the somatodendritic marker MAP2.

Similar to TDP-43, there is a similar debate in the field about whether the pathogenesis in FUS proteinopathy is mediated by a loss-of-function or a gain-of-toxicity mechanism (Gao et al., 2017; Ishigaki and Sobue, 2018; Sobue et al., 2018). On the one hand, mutant FUS leads to aberrant cytoplasmic localization and abnormal stress granule formation (Gao et al., 2017), and sequestration of essential spliceosome components (Sun et al., 2015; Yu et al., 2015). Further, increased FUS expression has been detected in FTLD cases (Deng et al., 2015). On the other hand, nuclear clearance of FUS and sequestration into aggregates leads to an impaired DNA damage response (Sama et al., 2014), impaired splicing (Ishigaki and Sobue, 2018; Lagier-

Tourenne et al., 2012), and decreased stability of synaptic-related transcripts (Udagawa et al., 2015).

With regard to downstream pathways affected by FUS, it was mentioned above that FUS appears to have a role in synaptic development and maintenance (reviewed in (Ling, 2018)). Besides direct interaction with spliceosome components, FUS was also found to regulate introns processed by the U11/U12 minor spliceosome (Reber et al., 2016). Similar to TDP-43, FUS has also recently been implicated in miRNA biogenesis and mitochondrial function, which will be summarized next.

1.3.1.c The role of FUS in miRNA biogenesis

Multiple recent studies have demonstrated that, like TDP-43, FUS also has a role in miRNA biogenesis (Eitan and Hornstein, 2016; Gascon and Gao, 2014). FUS was also discovered as an interacting subunit of the microprocessor complex (Gregory et al., 2004). FUS was found to be involved in a co-transcriptional recruitment of Drosha to sites of pri-miRNA transcription (Morlando et al., 2012), and can directly bind both miRNA precursors (Fernandes, 2012; Morlando et al., 2012) and mature miRNAs (Zhang et al., 2018). As an example, FUS regulates the expression of two miRNAs (miR-141 and miR-200a) that target FUS itself; this occurs via a feedforward loop involving the inhibition of Zeb1, a transcriptional repressor of these miRNAs (Dini Modigliani et al., 2014). Finally, a recent study demonstrated that FUS is critical for the proper silencing activity of multiple miRNAs (Zhang et al., 2018). FUS directly interacts with RISC component Ago2, mediated by the second RGG domain. Disrupting this interaction either by deletion of the RGG domain or by FUS mutation led to reduced silencing activity, independent of miRNA biogenesis. Finally, FUS-associated miRNAs have already by linked to neurodegeneration (Eitan and Hornstein, 2016; Gascon and Gao, 2014).

1.3.1.d The role of FUS in mitochondrial function

There have been mentions of a possible connection between FUS and mitochondria in previous work. An ultrastructure study of motor neurons in ALS-FUS patients revealed aggregations of ER and mitochondria near FUS inclusion bodies (Huang et al., 2010). Transgenic rats expressing a different ALS-associated R521C FUS had ubiquitin-positive inclusions that did not stain for FUS but did stain positive for COXIV, a component of complex IV (Huang et al., 2011). Cultured mouse motor-neurons transfected with other ALS mutants had smaller mitochondria in their axons (Tradewell et al., 2011). A comparison of wild-type and mutant FUS indicated Finally, similar to TDP-43, FUS was also found to regulate the VAPB-PTPIP51 interaction mediating ER-mitochondria communication (Lau et al., 2018; Stoica et al., 2016).

Additional evidence, in both non-neuronal and neuronal model systems, indicate that FUS binds to and regulates mitochondria-associated genes. One study looked at FUS binding in HEK293 cells, both wild-type and ALS mutants. An over-representation analysis identified "mitochondrion" and "mitochondrial inner membrane" as two pathways that were affected by mutant FUS binding as compared to wild-type (see Supplementary Table 3 from (Hoell et al., 2011)). A similar study with **[[FIX THIS]]** found that a different FUS mutation associated with multiple metabolic enzymes (Wang et al., 2015a). Another study looked at the interaction of FUS and RNA Polymerase II in HEK293 cells, showing a role for FUS in the phosphorylation of the C-terminal domain, regulating polyadenylation. Finally, an overrepresentation analysis of genes with at least 20% fold-change and significant p-value identified multiple mitochondrial gene ontology terms as hits (Schwartz et al., 2012).

Our group has recently demonstrated mitochondrial localization of FUS and discovered interaction of FUS with mitochondrial chaperonin HSP60 (Deng et al., 2015). Overexpression of wild-type or mutant FUS led to mitochondrial localization of FUS and mitochondrial fragmentation in multiple models, as seen in immunofluorescence and immuno-EM studies; a similar phenotype observed in post-mortem brain samples of FTLD-FUS cases. Along with the structural changes, we also observed a decrease in mitochondrial membrane potential, oxygen consumption rates, and ATP production (Deng et al., 2015). We further demonstrated that FUS-toxicity was mediated by HSP-60, with HSP60 knockdown partially rescuing the mitochondrial deficits. In a more recent study, we observed that wild-type or mutant FUS overexpression causes a dysregulation of PINK1 and Parkin protein levels, and also aberrant ubiquitination of Miro1, with a resulting deficit in mitochondrial retrograde movement in fly and mammalian neurons (Chen et al., 2016). Finally, we most recently have shown a direct interaction between FUS and members of complex V, especially ATP5B (Deng et al., 2018).

1.4 High-throughput Studies of TDP-43 and FUS

1.4.1 The Current Need for High-throughput Studies

Cells and tissues work as a unified whole for the sake of the organism, and thus it is an essential task for biologists to understand how different processes are integrated to produce a response, and how such integration is disrupted in diseases (Talbott, 2010; 2011). This is no less true for understanding the pathogenesis of complex diseases like ALS and FTD (Cooper-Knock et al., 2017; Ferrari et al., 2016; Fontana et al., 2015; Mao et al., 2017), nor for understanding the function of particular RNA-binding proteins like TDP-43 and FUS (Gama-Carvalho et al., 2017). This is especially so given all of the evidence to suggest that these proteins coordinate multiple steps of RNA processing from transcription to translation (as described in the previous section).

High-throughput studies can assay changes on a genome-wide scale ("omics" scale), allowing researchers the possibility of understanding this integration (Hawkins et al., 2010).

It is no surprise then that many high-throughput studies have already been done on TDP-43 and FUS. In fact, there have been 55 high-throughput datasets generated to study TDP-43, and 58 datasets to study FUS (including one of our own). The following section will summarize some of the studies, with a particular emphasis on RNA-Sequencing studies, the subject of Chapters 3 and 4.

1.4.2 Global Patterns of regulation by TDP-43 and FUS

High-throughput studies on FUS and TDP-43 have examined various roles of these proteins across the entire spectrum of RNA processing (see section 1.3 for details). In general, studies have found that there appears to be a common set of targets regulated by both proteins (Honda et al., 2014; Lagier-Tourenne et al., 2012), though there are also major differences in other targets or in competing regulation for a common target (Colombrita et al., 2015; Lagier-Tourenne et al., 2012). There is evidence to suggest that these proteins are connected in function (Ratti and Buratti, 2016). Interestingly, though both proteins target 1000s of transcripts in the genome, studies have concluded that there are only a modest number of changes (Colombrita et al., 2015; Polymenidou et al., 2011; Sama et al., 2014). This is an intriguing discrepancy.

1.4.3 Lack of overlap between TDP-43 studies

A small meta-analysis was done using three datasets generated from TDP-43 model systems (Buratti et al., 2013). They examined the overlap of hits predicted from three microarray studies and found zero hits that were identified in all three studies. Further, when examining the overlap of target genes affected by TDP-43 knockdown in non-neuronal cell lines, and targets bound by TDP-43 in the brain, they also found very little overlap. Other work has also found it

challenging to reproduce differential expression results from similar experiments (Subramanian et al., 2005). Work in both microarrays and in RNA-Seq has found that, independent of biological differences, the issue of low overlap is strongly dependent technical factors such as the RNA quality (Fasold and Binder, 2014), the platform and protocol used (Baran-Gale et al., 2015; Lahens et al., 2014), the depth of sequencing in RNA-Seq studies (Hart et al., 2013; Zhao et al., 2014), and the choice of annotation and pipeline (Baruzzo et al., 2017; Williams et al., 2016; Zhao and Zhang, 2015). Re-analyzing datasets can only address in a limited way the technical factors due to the experiment; it can, however, address the variation due to the choice of analysis pipeline but analyzing all of the data on the same pipeline.

1.4.6 Important functions missed by previous high-throughput studies

Re-analysis of previously generated data serves an important role in discovering new functions missed by past work. Focusing on just studies related to FUS and TDP-43, there have been new functions discovered through re-analysis of past data. For example, a study discovered that FUS has a role in minor intron splicing, and found the same pattern of regulation in a previous dataset that was missed by the original authors (Reber et al., 2016). Another study discovered a role for TDP-43 in suppressing cryptic exons, and also observed the same pattern of regulation in previous studies that were missed by the previous authors (Ling et al., 2015).

When it comes to analyzing specific pathways regulated in RNA-Sequencing or Microarray studies, recent work has demonstrated that there are important biases that impact the analysis (Timmons et al., 2015; Young et al., 2010). In particular, one has to be careful with length bias in RNA-Seq data (gene sets with longer genes tend to be selected for enrichment) when using counts (Young et al., 2010), and the results from an enrichment analysis are incredibly sensitive to the choice of the “background gene list” to use for comparison, with the

default background list provided by popular tools like DAVID or GeneOntology Enrichment are typically not suitable and will yield biased or uninformative results(Timmons et al., 2015).

1.5 Motivation and overview of this thesis

Given the above background, we hypothesized that FUS and TDP-43 both have roles in regulating the maintenance and function of the nervous system through their regulation of miRNA biogenesis, and through the regulation of mitochondria. To study this hypothesis, this work had three basic aims:

Aim 1: Systematically study the miRNA-mRNA network regulated by TDP-43 and FUS. Before this work, no systematic analysis of FUS and TDP-43 regulated miRNAs had been done. Any transcriptome-wide study was done using microarrays, which are limited by the annotation at the time. Further, no one had studied the network of interactions between FUS, miRNAs, and the target mRNAs, and how that network may impact critical processes in the nervous system. This aim was explored through the work described in Chapter 2.

Aim 2: Develop a pipeline to analyze RNA-Sequencing Data related to FUS and TDP-43 model systems. Initial failures with qPCR validation (see **Figure 4.7**) and reflection on the nature of RNA-Sequencing data required the need to design a new method for normalizing the data, as well as a new simulation protocol to properly benchmark RNA-Seq tools. This led to the development of **sleuth-ALR** and **absSimSeq**. Chapter 3 describes this work.

Aim 3: Using both molecular and bioinformatic methods, evaluate the role of FUS in mitochondrial function. Despite evidence of mitochondrial regulation, no high-throughput study had systematically studied putative FUS- and TDP-43-regulated mitochondrial genes. Our FUS overexpression work motivated us to study FUS KO impact on mitochondria. Unpublished

observations motivated us to pursue further studies in HEK FUS KO cells. Chapter 4 describes some initial work to achieve this goal.

Preface to Chapter 2

The work in this chapter appeared in the following publications:

Chen, X., Fan, Z., McGee, W., Chen, M., Kong, R., Wen, P., Xiao, T., Chen, X., Liu, J., Zhu, L., et al. (2018). TDP-43 regulates cancer-associated microRNAs. Protein & Cell 9, 848–866.

Xiaowen Chen and Zhen Fan need to be acknowledged for the work that identified putative TDP-43-regulated miRNAs. Mengmeng Chen also needs to be acknowledged for conducting the cell migration and TDP-43 binding studies that appears in **Figure 2.4**.

Yonggui Fu and Anlong Xu need to be acknowledged for generating the paired mRNA-Seq and small-RNA-Seq dataset from the FUS KO mice.

Chapter 2: miRNA-mRNA networks regulated by TDP-43 and FUS in cancer and in the nervous system

2.1 Introduction

2.1.1 Evidence that TDP-43 may have a role in cancer mediated by miRNAs

As discussed in the introductory chapter, both TDP-43 and FUS have important roles to play in microRNA biogenesis. Before this work, however, there had been few systematic studies of miRNAs regulated by TDP-43 and FUS, and what had been done used microarrays. However, compared with microarrays, high-throughput may be a better choice. The probes on a microarray are limited by the annotations available at the time of design, whereas sequencing is not tied to an annotation and therefore can detected unannotated sequences; there are also known issues with cross-hybridization (Leshkowitz et al., 2013; Mestdagh et al., 2014). We thus set out to assay microRNAs regulated by TDP-43 and FUS using small-RNA-Sequencing. For TDP-43, we conducted a knockdown experiment using a pool of siRNAs in three cell lines (SH-SY5Y, SNB19, and HT-22), and performed small-RNA-Sequencing (Chen et al., 2018). In this study, despite conducting the work in neuronal-like cell lines, we identified several TDP-43-regulated miRNAs that had known roles in cancer pathogenesis (Chen et al., 2018). This prompted us to explore a possible role for TDP-43 in cancer through its regulation of miRNAs.

There is ample evidence to suggest a connection between TDP-43 and FUS proteinopathies and cancer. FUS was originally discovered in the context of cancer (Crozat et al., 1993; Rabbitts et al., 1993). Recent work has also implicated TDP-43 in cancer (Campos-Melo et al.). In particular, a variant near the TARDBP gene (encoding TDP-43) was linked to susceptibility of Ewing Sarcoma (Postel-Vinay et al., 2012). TDP-43 was identified as

differentially expressed after a breast cancer cell line, MCF-7, was treated with the known anti-cancer agent curcumin (Fang et al., 2011). TDP-43 was also shown to directly regulate glycolysis via a miRNA in a hepatocellular carcinoma cell line (Park et al., 2013). Finally, multiple studies have shown an inverse association the risk for cancer and for neurodegeneration (Gibson et al., 2016; Katisko et al., 2018; Umansky, 2018); a few have presented evidence that this inverse relationship may be explained by the role of miRNAs in both (Murmann et al., 2018; Umansky, 2018). It thus seems reasonable to hypothesize that TDP-43 has a role in cancer pathogenesis and that this role may be mediated by TDP-43-regulated miRNAs. Because a majority of the identified TDP-43-regulated miRNAs with known roles in cancer were involved in lung cancer, we decided to look more closely at lung cancer.

2.1.2 The need to use a network approach for studying miRNA-mRNA interactions

How do miRNAs regulate their target mRNAs? Canonically, RISC, with a loaded mature miRNA, binds a target mRNA and targets it either for translational inhibition or for degradation (Wilczynska and Bushell, 2015). Most miRNA-mRNA interactions result in only modest reduction of their target gene. There are two major functional consequences to the target gene as described by Bartel and Chen: either an mRNA that should not be present is repressed completely (e.g. a glial-specific gene in the context of a mature neuron), or a gene that requires a specific range of protein can be tuned to the right level by the miRNA (Bartel and Chen, 2004). They used the analogy of a dimmer switch or a rheostat, and called the former case an “off switch” interaction, and the latter case a “tuning” interaction. An example of the “off switch” interaction in neurons is mir-9 shutting off onecut1 (OC1) during transition from early-born to late-born motor neurons in chicks (Luxenhofer et al., 2014). An example of the “tuning” interaction in neurons can be seen with miR-8 and atrophin in Drosophila (Karres et al., 2007).

In many cases, miRNAs are regulated by the same transcription factors (TFs) as their targets, and this facilitates their ability to facilitate an off switch or a tuning interaction (Le et al., 2015; Osella et al., 2011; Vera et al., 2012). In a coherent feedforward loop, a TF shuts off the expression of a gene and activates a miRNA, which targets that gene; this facilitates an "off switch" interaction where the miRNA acts a fail-safe to prevent any mRNA transcripts that leak through the TF-mediated repression from being translated. In an incoherent feedforward loop, a TF activates the expression of a gene and a miRNA which targets that gene; this facilitates a "tuning" interaction where the miRNA dampens fluctuations in the target gene and maintains it in a specific range (Osella et al., 2011). Of note, miRNAs and mRNAs interacting in an "off switch" capacity tend to have anti-correlation, and whereas those interacting a "tuning" capacity tend to be correlated (Osella et al., 2011).

2.1.3 Aims of this study

With the above as background, we had two aims. The first aim was to understand the network of TDP-43-regulated miRNAs and their downstream target mRNAs, as well as the processes regulated by these miRNA-mRNA interactions. To do this, we had to design a novel pipeline to generate a network of predicted causal miRNA-pathway associations mediated by miRNA-mRNA interactions (see Methods and **Figure 2.1**). We applied this pipeline to the Cancer Genome Atlas data available on two lung cancers: lung adenocarcinoma (LUAD) and lung squamous carcinoma (LUSC). This pipeline identified several TDP-43-regulated miRNAs with putative roles in cancer pathogenesis. One of these miRNAs, miR-423-3p, was predicted to have a role in cell migration, mediated through four downstream target genes. We therefore did follow-up work to demonstrate that TDP-43 directly bound miR-423-3p in a lung cancer line,

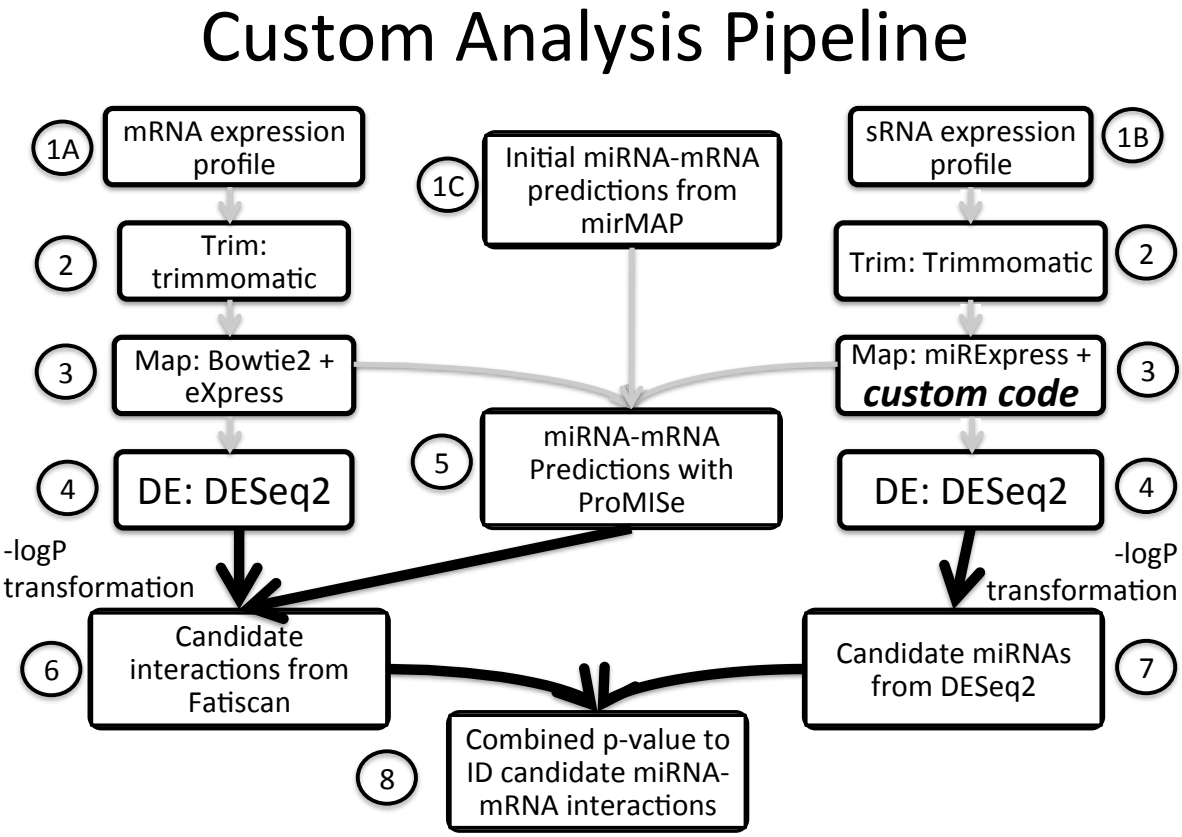

***Figure 2.1:* Flowchart of analysis pipeline.** *We designed a custom pipeline to analyze our data to identify candidate miRNA and miRNA-mRNA interactions that were predicted to be influencing cellular processes important for neuronal function. Step 1: We start with three inputs (shown across the top): paired expression profiles from (A) mRNA-enriched RNA-Seq and (B) small-RNA-Seq (to asses miRNA expression levels), and (C) initial target site predictions from mirMAP (2013/01 pre-calculated predictions). Step 2: For the expression profiles, we preprocessed the fragments using trimmomatic to remove adapter contamination and low-quality and small reads. Step 3: We then mapped these reads to the mm10/GRCm38 mouse genome using Ensembl annotation (v78). For mRNA-Seq, we used bowtie2 streamed into eXpress, allowing multiple mappings as possible and utilizing all of them; for small-RNA-Seq, we used miRExpress and then custom code to prepare the data for differential expression analysis. Step 4: We then submitted the effective counts from both respectively to DESeq2 for differential expression analysis. Step 5: Using the effective counts from both, along with the mirMAP target site predictions, we calculated miRNA-mRNA interaction predictions for each sample individually using ProMISe; we kept only predictions made with non-zero probability in all samples analyzed. Step 6: annotating each transcript with the miRNAs predicted to target them (as calculated by ProMISe), we submitted all transcripts ranked by their –log(adjusted p-value) * sign of their change (+ for up; - for down) to Fatiscan to look for enrichment of differentially expressed genes targeted by miRNAs. Step 7 and 8: We took the adjusted p-values from DESeq2 for each miRNA and used our modified SPIA analysis to identify candidate miRNAs which were both differentially expressed (determined by DESeq2) and had enrichment among its predicted targets for DE genes (determined by Fatiscan).*

and that TDP-43 promoted cell migration via a miR-423-3p-dependent mechanism. Thus, our

network analysis pipeline produced a prediction that was experimentally validated.

The second aim was to apply this pipeline to the role of miRNAs in the nervous system. This pipeline requires paired mRNA and miRNA expression profiles to directly predict interactions between miRNAs and their target mRNAs. Since no such dataset had been generated for either TDP-43 or FUS, we generated a dataset of paired mRNA and miRNA expression profiles from the brains of FUS KO mice and their wild-type littermates. Using this dataset, we generated a network of miRNA-pathway associations mediated by predicted miRNA-mRNA interactions. This predicted a role for FUS-regulated miRNAs in synaptic maintenance and regulation, especially calcium signaling.

2.2 Results

2.2.1 TDP-43-regulated miRNAs are predicted to influence multiple pathways in lung cancer

In order to predict how TDP-43 might be involved in lung cancer via the miRNAs that it regulates, we designed an analysis pipeline that combined ProMISe (probabilistic miRNA-mRNA interaction signature) (Li et al., 2014), DESeq2 (Love et al., 2014), Fatiscan (Al-Shahrour et al., 2007b) and FatiGO (Al-Shahrour et al., 2007a). See the Methods (Section 2.3.2) and **Figure 2.1** for a graphic of the analysis pipeline. For this analysis, we focused on the two datasets studying NSCLC, lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD), with samples that had paired miRNA-mRNA expression profiles available from The Cancer Genome Atlas as of July 2014 (Collins and Barker, 2007).

First, the miRNAs were tested for differential expression using DESeq2; out of the 1100 human miRNAs in the miRBase v21 database (Kozomara and Griffiths-Jones, 2014), this resulted in 417 and 563 miRNAs being differentially expressed in LUAD and LUSC samples, respectively, versus control samples. MiRNAs putatively regulated by TDP-43 were over-

represented in these groups (hypergeometric test p-value for LUAD (57/417 vs 83/1033) and LUSC (61/563 vs 83/1037): 5.24e-8 and 1.48e-4, respectively). Because TDP-43 may not regulate all of these miRNAs in this context, we examined the correlation between each miRNA and TDP-43 expression in lung cancer samples. We performed Pearson correlation, and after correcting for multiple hypothesis testing, found 408 and 467 miRNAs significantly correlated with TDP-43 in LUAD and LUSC samples, respectively (FDR<0.1). MiRNAs putatively regulated by TDP-43 had a trend for overrepresentation in these groups as well (hypergeometric test p-value for LUAD (39/408 vs 83/1033) and LUSC (45/467 vs 83/1037): 0.091 and 0.059, respectively).

To identify which miRNAs had an enrichment for differentially expressed targets, we used the combination of ProMISe and Fatiscan. From the ProMISe step, out of the 1033 and 1037 miRNAs in LUAD and LUSC samples that had some expression, there were 213 and 274 miRNAs that had at least 5 predicted targets in every LUAD and LUSC sample, respectively; miRNAs regulated by TDP-43 were also over-represented in these groups (hypergeometric test p-value for LUAD (67/213 vs 83/1033) and LUSC (74/274 vs 83/1037): 3.36e-35 and 1.89e-36, respectively). A ranked list of the transcripts (using log fold change and log differential expression p-value from DESeq2) was submitted along with the predicted miRNA-mRNA interactions from ProMISe as custom annotations to Fatiscan, as part of the Babelomics v4 suite (Medina et al., 2010).

In order to identify miRNAs that were both differentially expressed and had enrichment for differentially expressed targets, we then combined the Fatiscan results with earlier DESeq2 results for the miRNAs to get a joint p-value. We applied filter criteria to reduce the list of

miRNAs to the most relevant one (for details, see Methods, Section 2.3.2.d). The results of this step for LUAD and LUSC are shown in **Table A1.1**.

To determine what biological processes were related to the identified targets, we extracted the unique transcripts from the previous step and submitted them to FatiGO, also part of the Babelomics v4 suite; this tool performs an overrepresentation analysis for gene ontology and pathway terms. Among the down-regulated transcripts, the most significant hits included integrin cell surface interactions and negative regulation of cell proliferation; among the up-regulated transcripts, the most significant hits included nucleotide synthesis, cell cycle checkpoints, and RNA processing. This suggests that TDP-43–regulated miRNAs may play a role in promoting carcinogenesis and metastasis. The full list of hits can be found in **Table A1.2**.

From this analysis pipeline, we defined "predicted causal interactions" as those miRNA-mRNA interactions between putative TDP-43-regulated miRNAs and target mRNA transcripts with annotations in the processes discovered by FatiGO. **Figure 2.2A-B** shows the representative network graph of up-regulated miRNAs and down-regulated transcripts in LUAD (one network of 7 up-regulated miRNAs, 50 down-regulated transcripts, and 13 processes; another network of 4 down-regulated miRNAs, 62 transcripts, and 17 processes), and **Figure 2.2C-D** for the LUSC network, which was much larger. See **Table A1.3** for the full node and edge lists.

In summary, our analysis pipeline identified a number of putative TDP-43-regulated miRNAs which target several transcripts that have roles in cancer biology. One of these was experimentally examined further for its role in lung cancer.

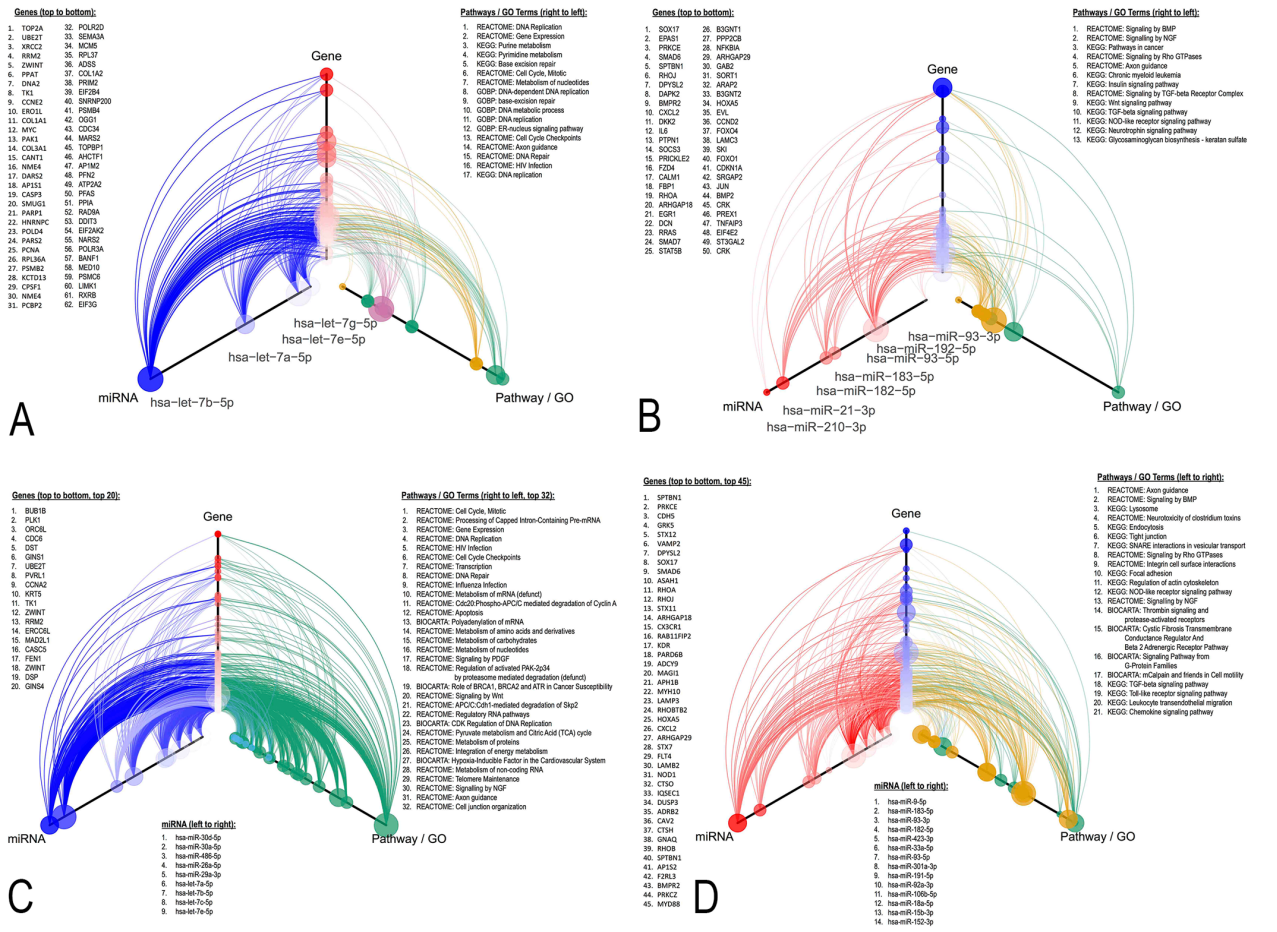

***Figure 2.2:* The network graph of miRNA-mRNA interactions and the significant FatiGO terms associated with the target mRNAs in LUAD and LUSC samples.** *These are hive plots linking miRNAs to putative mRNA targets, and gene ontology terms identified to be significantly enriched in these diseases using Fatiscan. The radial distance is the rank of the nodes within their respective groups. The node size is proportional to the number of connections for a node. The color of the miRNA and gene nodes is related to the statistical significance of differential expression, and the color of the Pathway / GO term is related to the category. Finally, the color of the miRNA-mRNA connections is related to the statistical significance of the Fatiscan step, and the color of the mRNA-FatiGO connection is also related to the category. Note that some of the significant pathway terms are omitted for clarity. See Tables S1.2 and S1.3 for the full list of significant pathways and the list of edges and nodes depicted. (A) down-regulated miRNAs targeting up-regulated transcripts in LUAD samples. (B) Up-regulated miRNAs targeting down-regulated transcripts in LUAD samples. (C) Down-regulated miRNAs targeting up-regulated transcripts in LUSC samples. (D) Up-regulated miRNAs targeting down-regulated transcripts in LUSC samples.*

2.2.2 TDP-43 associated miR-423-3p promotes lung cancer cell migration

Other work performed in our lab identified miR-423-3p as a putative TDP-43-regulated miRNA; we also performed additional work to show that TDP-43 directly bound miR-423-3p in SH-SY5Y cells (Chen et al., 2018). From our analysis pipeline, miR-423-3p was one miRNA

that met all of our criteria in LUSC samples: it was differentially expressed, significantly correlated with TDP-43, and had targets with statistically overrepresented pathway annotations (**Figure 2.2D**). Of the four mRNA targets that were hits, three (CRK, LCP2, and ITGA9) were related to the reactome pathway "Integrin Cell Surface Interactions"; even when including the rest of the differentially expressed targets of miR-423-3p, this was the only significant pathway identified by FatiGO (data not shown). Thus, we hypothesized that TDP-43 might influence lung cancer cell migration via miR-423-3p.

To test this hypothesis, we performed TDP-43 knockdown on H1299 lung cancer cells and measured cell migration using the transwell migration assay. Transwell migration assays using human lung cancer H1299 cells showed statistically significant reduction in cell migration by TDP-43 knockdown (**Figure 2.3A-C**). In order to address whether this inhibition of cell migration was related to TDP-43-regulated miRNAs, we co-transfected H1299 cells with TDP-43 siRNAs and one other TDP-43-regulated miRNA previously identified with lung cancer, miR-146b-5p. MiR-146b-5p was selected as a negative control because it had already been shown not to affect cell migration in a different lung cancer cell line (A549) (Patnaik et al., 2011). After co-transfection with miR-423-3p, cell migration increased significantly (p-value<0.05) as compared with cells transfected with TDP-43 siRNAs alone (**Figure 2.3A-B**). Co-transfection with miR-146b-5p did not rescue cell migration (**Figure 2.3A-B**). Similar to results from SH-SY5Y cells (Chen et al., 2018), examination of the interaction between miR-423-3p and TDP-43 using RNA immunoprecipitation (RIP) and RNA pull-down assay showed that miR-423-3p interacts with TDP-43 in H1299 lung cancer cells (**Figure 2.3D-E**). Thus, these results suggest that TDP-43 promotes lung cancer cell migration through the direct regulation of

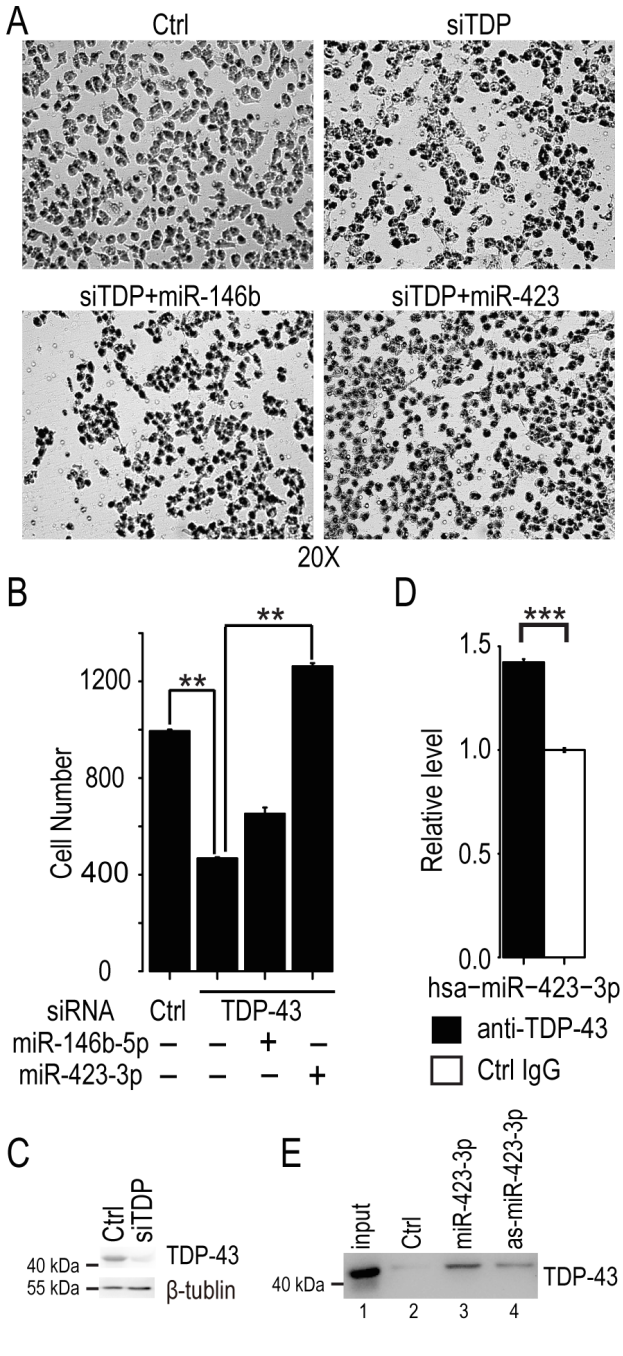

Figure 2.3: Effect of TDP-43 regulated miRNAs on lung cancer cell migration. *(NOTE: this is Figure 5 from Chen X, Fan Z et al 2018)* ***(A)*** *H1299 cell migration after transfection with either control siRNAs or TDP-43 siRNAs alone (upper panel) or in combination with has-miR-146b-5p (bottom left) or has-miR-423-3p (bottom right).* ***(B)*** *Quantification of migrated cells.* ***(C)*** *Western blot showing the effect of TDP-43 knockdown in H1299 cells.* ***(D)*** *RIP coupled qRT-PCR assay of interaction between TDP-43 and has-miR-423 in H1299 cells. Enrichment was determined as miR-423 associated to TDP-43 IP relative to control IgG.* ***(E)*** *RNA pull-down assays of the interaction between has-miR-423 and TDP-43 (combined with qRT-PCR) in H1299 cells. Lane 1, ~3% input; Lane2, negative control; Lane 3 and 4, biotinylated has-miR-423-3p and antisense-hsa-miR-423-3p, respectively. (n = 3; means ± SEM; ** = p < 0.05; *** = p < 0.01.)*

miR-423-3p, corroborating the prediction from the functional annotation pipeline that TDP-43 is a tumor promoter.

2.2.3 FUS-regulated miRNAs in the brain are predicted to regulate synaptic and calcium signaling pathways

After the success of the pipeline on cancer data, we sought to generate predictions from miRNAs regulated in the nervous system. To do this, we generated paired mRNA-Seq and small-RNA-Seq expression profiles from the brains of FUS KO mice and compared them to the profiles from the brain of wild-type littermates. Out of roughly 1000 mature miRNAs that were expressed in the brain (average normalized read count ≥ 5), 17 miRNAs were down-regulated and 15 were up-regulated after FUS KO (DESeq2 FDR≤0.1). Out of roughly 53,000 transcripts that were expressed (also

average normalized read count ≥5), 1035 transcripts were down-regulated and 627 were up-regulated after FUS KO (DESeq2 FDR≤0.1). The full results for miRNAs and mRNAs are available in **Table A1.4**.

To generate predictions of FUS-regulated miRNAs regulating downstream targets and pathways, we used a similar pipeline as above to get predicted miRNA-mRNA interactions; we then used Fatiscan to identify which miRNAs had an enrichment of up- and/or down-regulated transcripts. Fatiscan enrichment of targets and miRNA differential expression were then combined into one p-value. 13 of the 17 down-regulated miRNAs and 13 of the 15 up-regulated miRNAs also had an enrichment of down-regulated predicted targets (Fatiscan FDR≤0.05). In stark contrast, no differentially expressed miRNAs had an enrichment for up-regulated predicted targets. The full list of combined results is available in **Table A1.5**.

We then looked at pathways enriched among the identified targets using FatiGO and constructed a network of predicted miRNA-mRNA causal interactions and mRNA-pathway associations. When we further restricted the interactions to mRNA targets with at least one GO term identified as significantly enriched, miRNA-mRNA interactions that were conserved in humans, and miRNAs with DESeq2 differential expression FDR<0.1, we were left with the small network shown in **Figure 2.4**. The reduced network including pathway terms can be seen in **Figure A1.1**. Importantly, many of the GO terms identified related to synaptic maintenance and function. In summary, the miRNA-mRNA predicted causal network and functional annotation pipeline revealed putative FUS-regulated miRNAs with predicted roles in regulating synaptic function.

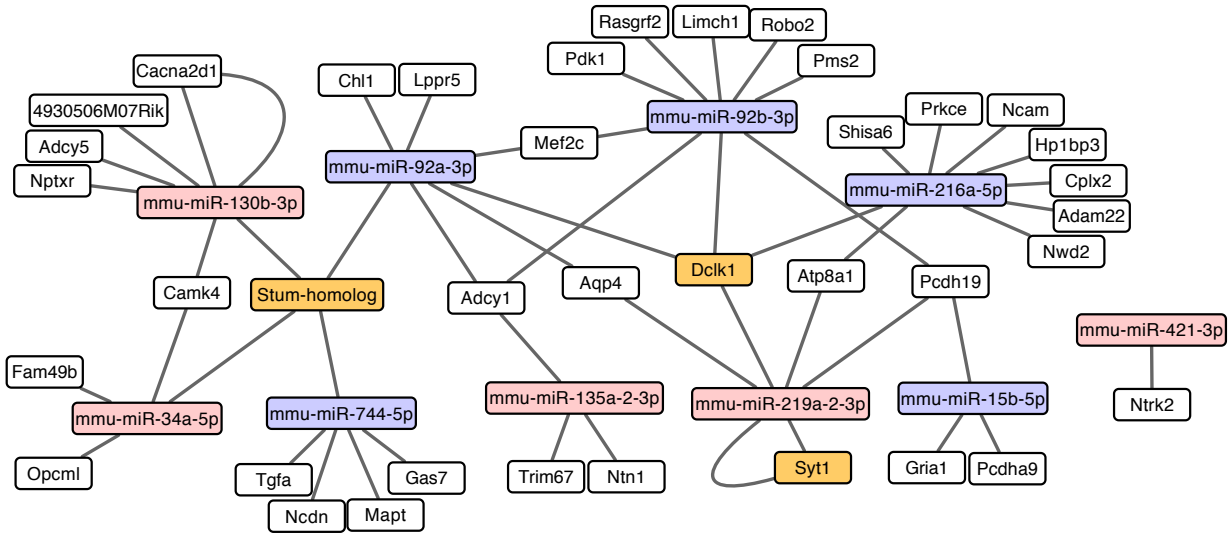

***Figure 2.4:* Network graph of miRNA-mRNA interactions in FUS KO whole mouse brains compared to wild-type.** *The above graph shows all significant miRNA-mRNA interactions that have evidence of conservation between mouse and human. All miRNAs shown are up-regulated in the FUS KO condition, suggesting that FUS inhibits their expression. The miRNAs in red have a differential expression FDR <0.05, and the miRNAs in purple have an FDR <0.1 (as estimated by DESeq2). Each edge between a miRNA and target mRNA represents one transcript predicted to be regulated by the miRNA; there are a few examples where the miRNA is predicted to regulate more than one transcript isoform of a gene, and this is a represented by multiple edges. The genes in orange are three genes that were interesting candidates for downstream analysis. This graph was constructed in Cytoscape.*

2.2.4 Follow-up Work

Because of changing circumstances in the lab and technical difficulties (data not shown), we did not pursue further work related to miRNAs regulated by TDP-43 or FUS. See Section 5.1 for discussion of further hypotheses to pursue in this area.

2.3 Methods

2.3.1 TCGA Data collection for miRNA-mRNA Functional Annotation and Predicted Causal Network

See **Figure 2.1** above for a summary of the analysis pipeline. From the Cancer Genome Atlas (TCGA) database for lung squamous cell carcinoma and lung adenocarcinoma samples (Collins and Barker, 2007), we extracted the data from all samples that had paired miRNA-Seq and RNA-SeqV2 profiles available as of July 2014 (330 tumor, 37 control for LUSC; 422 tumor,

19 control for LUAD). From the pre-calculated human target predictions from miRanda ((Betel et al., 2010); http://www.microrna.org/microrna/getDownloads.do), a matrix was generated using a python script reporting the number of binding sites for each miRNA-mRNA interaction in humans. Only predicted sites with a "good mirSVR" score were used, irrespective of conservation. A Perl script was then used to assign the TCGA raw miRNA counts (*.isoform.quantification.txt files) to the mature miRNAs, as defined by miRBase version 21 (Kozomara and Griffiths-Jones, 2014). Another Perl script was used to isolate the mRNA expression estimates (*.isoforms.normalized_results files) for the next steps.

2.3.2 miRNA-mRNA Functional Annotation and Predicted Causal Network Pipeline

2.3.2.a ProMISe analysis

Probabilistic **M**iRNA-mRNA **I**nteraction **S**ignatur**e** (ProMISe) is a recently developed technique (Li et al., 2014) that incorporates information about the number of binding sites a miRNA has on a target gene as well as expression levels of both the miRNAs and the target genes. Unique to ProMISe, though, is the generation of a competition model of miRNAs competing for a particular mRNA, and mRNAs competing to be inhibited by a particular miRNA. The joint model of these two competition models outperforms all other available miRNA-mRNA interaction prediction tools; it also has the advantage of predicting these interactions within a single sample (Li et al., 2014). For our data, the matrix from the miRanda predictions, the processed miRNA expression profiles, and the normalized mRNA isoform expression profiles were used as input for ProMISe, using the "joint model", to generate for each sample a "ProMISe signal" consisting of a probability matrix of any particular miRNA targeting any particular gene. From the ProMISe signature for each sample, all miRNA-mRNA interactions with non-zero probability were counted as predicted miRNAs targets for that

sample. For each miRNA, only interactions seen in all samples were included as a "predicted target" for downstream analyses. We restricted analysis to those miRNAs that had at least five targets.

2.3.2.b Differential expression analysis and ranking transcripts

The isoform counts for miRNAs and mRNAs were submitted to DESeq2 (Love et al., 2014) for differential expression analysis using the standard settings. For miRNAs, the raw aggregated counts for mature miRNAs were used. For mRNAs, the RSEM normalized estimated counts were used; this is analogous to using estimated transcript abundances, as described in a recent paper (Soneson et al., 2016). In order to rank the transcripts for the Fatiscan step, an "adjusted rank" was used to give the most weight to transcripts that had the most expression, the most log-fold change, and the most statistically significant change. If the transcript had a base mean of 30 or less, then its rank was log10 of its base mean plus the absolute value of its log2 fold change; otherwise, its rank was those two items plus the absolute value of log10 of its adjusted *p*-value. Then the rank was given the same sign as the transcripts' fold change (negative for down-regulated; positive for up-regulated).

2.3.2.c Fatiscan analysis

Fatiscan (Al-Shahrour et al., 2007b) is a tool that is threshold-independent, using a heuristic to define a partition of a ranked list of genes or transcripts to identify whether a set of them are overrepresented among the most up-regulated or most down-regulated. In our case, we submitted a list of custom annotations based on the ProMISe results, with each transcript annotated with the miRNAs that target them, as well as the "adjusted rank" list generated in the previous step. We then ran Fatiscan with the options "remove duplicates", "Fatiscan" model, "Two-tailed Fisher's

Exact Test", and our custom miRNA annotations as the database to test. The results were downloaded, and the adjusted p-values were extracted.

2.3.2.d Selecting candidate miRNAs

SPIA (Tarca et al., 2009) is a technique that combines two dimensions of data to estimate which pathways are significantly altered: overrepresentation of differentially expressed genes, and their own metric estimating the "pathway perturbation". The p-values for both can be combined using Fisher's product method, or their "normal inversion" method, which gives greater weight when both dimensions have a low p-value. In our study, we used the latter method of combining p-values to combine the DESeq2 adjusted *p*-value and the Fatiscan adjusted *p*-value for each miRNA. In this way, our goal was to identify a miRNA that is both perturbed and has a large number of perturbed downstream targets. This would lead us to predict that this miRNA is affecting the network through its targets. This combined *p*-value was then adjusted using the Benjamini-Hochberg method.

We identified all miRNAs that had an adjusted combined p-value < 0.05, and then applied four criteria to select candidate miRNAs: (A) the miRNA had to have a DESeq2 differential expression FDR < 0.1; (B) the targets of the miRNA had to be changing in the opposite direction (if the miRNA is up-regulated, the targets must be down-regulated, and vice versa); (C) from earlier work, the miRNA had to have a significant change in at least one of three cell lines (SH-SY5Y, SNB19, or HT22) after TDP-43 knockdown (Chen et al., 2018); (D) the expression profile of the TDP-43-regulated miRNA had to have a statistically significant correlation (FDR < 0.1) with the TDP-43 expression profile, suggesting that TDP-43 was regulating this miRNA in lung cancer.

To calculate the correlation, we extracted out the TDP-43 normalized gene counts from the TCGA data, and then performed a Pearson correlation of the TDP-43 gene counts against each miRNA's normalized counts (as calculated by DESeq2). We took the *p*-value of that correlation, and adjusted it using the Benjamini-Hochberg method. The resulting list of miRNAs for each combination of miRNA-mRNA interactions (down-regulated miRNAs targeting up-regulated mRNAs, and vice versa for LUAD and LUSC each) were submitted for the FatiGO step.

2.3.2.e FatiGO analysis

To generate a list of functional annotations, the transcripts identified extracted from the targets of each candidate miRNA. Four functional groups were tested separately: the down-regulated targets of up-regulated miRNAs and the up-regulated targets of down-regulated miRNAs for LUAD and LUSC each. The first step was to convert the UCSC IDs to gene names. A Perl script with the June 2011 TCGA human genome annotation (the annotation used at the time of data generation; available at https://www.synapse.org/#!Synapse:syn1356421), along with the current kgXref_table and the versions 5 and 6 from the UCSC database were used to construct a table converting the UCSC transcript IDs to gene names, with some manual updating of those names using the Ensembl and Unigene databases. The resulting lists were submitted as gene lists to FatiGO (Al-Shahrour et al., 2007a), as part of the Babelomics 4.3 suite ((Medina et al., 2010); v4.babelomics.org). Each gene list was compared against the human genome; the gene ontology biological process, gene ontology molecular function, BIOCARTA, KEGG, and Reactome databases were tested using the default settings.

2.3.2.f Construction of a predicted causal interaction network

From all of the above results, a network of predicted causal links, from TDP-43 to lung cancer through TDP-43-regulated miRNAs and their targets, was constructed based on the significant targets that had at least one annotation. The resulting interaction network graphic was constructed using the HiveR R package (http://academic.depauw.edu/~hanson/HiveR/HiveR.html), based on the principles of the Hive Plot (Krzywinski et al., 2012). A python script was used to convert the various attributes (e.g., rank) to hive plot characteristics (e.g. node color). Each item was treated as a node on one of three axes: miRNAs, mRNAs, and pathway terms. The miRNA-mRNA edges are significant interactions identified by our pipeline; the mRNA-term edges are significant annotations identified by the FatiGO step. The rank of the node was mapped to the radial distance; the signed log10 of the FDR was mapped to the color for the miRNA and mRNA nodes; the signed log10 of the Fatiscan result was mapped to the miRNA-mRNA edges; the database category was mapped to the pathway term node color and to the mRNA-term edges; finally, the number of connections was mapped to the size of each node.

2.3.3 Culture of H1299 Cells and Transwell Migration Assay

Human non-small cell lung cancer cells (H1299) were cultured in DMEM supplemented with 10% fetal bovine serum at 5% CO2 and 37°C. For the *in vitro* cell migration assay, 5×10^4 cells were suspended in 0.5 mL DMEM without serum, and then plated into the transwell inserts (BD Biosciences). 0.75 mL DMEM with serum was added to the bottom well. Cells were incubated for 12 hours, fixed in 75% ethanol for 10 minutes, and stained by crystal violet for 30 minutes. Cells that migrated cross the membrane were counted under a microscope from 6 randomly selected fields (at 20× magnification).

2.3.4 Generation of FUS KO Brain paired mRNA-Seq/small-RNA-Seq dataset

This experiment was done using the Hicks FUS KO model (Hicks et al., 2000). Brains were collected from E18-E19 FUS knockout embryos or from their littermate wild-type controls. Total RNA was extracted using Trizol (Invitrogen). Two aliquots were taken from each sample to generate paired libraries for mRNA-Sequencing and small-RNA-Sequencing. The libraries were prepared and sequenced by RiboBio (Guangzhou, China; en.ribobio.com/siteen/).

The mRNA-Sequencing libraries were prepared using the TruSeq® Stranded mRNA HT Sample Prep Kit (Illumina). Briefly, poly-adenylated RNA was isolated and then fragmented to 200 nucleotides. The sequences used for the trimming of the mRNA libraries were the following: 5'adapter (3' end of i5 adapter): 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'; 3'adapter (5' end of i7 adapter): 5'-GATCGGAAGAGCACACGTCTGAACTCCAGTC-3'.

The small-RNA-Sequencing libraries were prepared using the TruSeq® Small RNA Sample Prep Kit (Illumina). The 5'adapter used for trimming these libraries had the sequence 5'-GUUCAGAGUUCUACAGUCCGACGAUC-3', and the 3'adapter used for trimming had the sequence 5'-AGATCGGAAGAGCACACGTCT-3'.

Following cDNA synthesis and PCR amplification, cDNA library quality was assessed using an Agilent 2200 TapeStation and Qubit 2.0. Both sets of libraries were sequenced using an Illumina Hiseq 2500. For the mRNA-Seq data, 101-nucleotide paired-end reads were generated; for the small-RNA-Seq data, 51-nucleotide single-end reads were generated.

2.3.5 Estimates of FUS KO Brain mRNA and miRNA abundances

The raw reads from the mRNA-Seq data were first trimmed using trimmomatic (Bolger et al., 2014). For trimmomatic, the parameters "ILLUMINACLIP:adapters.fa:2:30:10 LEADING:4 TRAILING:4 MINLEN:36" were used. The trimmed reads were then mapped to

the Ensembl mouse transcriptome (GRCm38, Ensembl version 77) using bowtie2 (Ben Langmead and Salzberg, 2012) and eXpress (Roberts and Pachter, 2012). The bowtie2 command was “bowtie -aS -X 800 -q --no-mixed”. The estimates from eXpress were used directly for differential expression.

The raw reads from the small-RNA-Seq data were first trimmed using trimmomatic. The trimmomatic parameters used were “ILLUMINACLIP:adapters.fasta:2:30:6 LEADING:4 TRAILING:4 MINLEN:10” with the adapter sequences described above. They were then mapped to the miRNA sequences from miRBase version 21 using miRExpress 2.1.4 (Wang et al., 2009). The datasets were downloaded from the miRExpress website (mirexpress.mbc.nctu.edu.tw). Alignment identity of 85% was used to account for isomiR variation (Neilsen et al., 2012).

A custom python script was used to take the “alignment” files output by miRExpress and calculate an alignment score using the Smith-Waterman algorithm (match = 3; mismatch = –2; indel = –7) and generate statistics about any isomiR variation. A second python script was then used to filter out bad alignments (miRExpress alignment score <20) and recalculate the total expression of a mature miRNA when reads mapped to multiple locations (assuming a naïve equal probability of an ambiguous read mapping to each miRNA). These adjusted read counts were used for differential expression.

Differential expression for both mRNAs and miRNAs was assessed using DESeq2 with default parameters (Love et al., 2014).

2.3.6 miRNA-mRNA Functional Annotation and Predicted Causal Network Pipeline for the FUS KO Brain dataset

The same pipeline described in section 2.3.2 above was used for the FUS KO Brain dataset with the following modifications. FUS-regulated miRNAs were defined as any miRNA that was identified as differentially expressed between FUS KO and control samples using DESeq2 with FDR<0.05. For the ProMISe analysis, we then used the pre-calculated predictions from mirMAP (Vejnar and Zdobnov, 2012), which predicts the percent of repression of an mRNA target by a miRNA; we defined any non-zero repression as a predicted target. Because of the large number of miRNAs after the SPIA combined analysis, we also restricted analysis to differentially expressed miRNAs that also had an Fatiscan FDR<0.05. For the final network, we restricted interactions to those where there was evidence of conservation between mice and humans, the gene had a GO term identified as significant.

2.3.7 Data availability

The code and steps needed to reproduce the TCGA functional annotation and predicted causal network pipeline can be found on Github: https://github.com/warrenmcg/TDP43_miRNA_Paper. The code for the FUS network predictions are available upon request.

Preface to Chapter 3

The work in this chapter will appear in the following:

McGee, W.A., Pimentel, H., Pachter, L., and Wu, J.Y. (2019). Compositional Data Analysis is necessary for simulating and analyzing RNA-Seq data.

Harold Pimentel and Lior Pachter need to be acknowledged for the design of the sleuth experiments highlighted in Figures 3.4 and 3.5, and Harold especially needs to be acknowledged for developing the initial codebase to run the simulation and benchmarks from which this work builds.

CHAPTER 3: Development and Benchmarking of Compositional Normalization for RNA-Sequencing Data

3.1 Introduction

High-throughput methods, including RNA-Seq, are frequently used to determine what features—genes, transcripts, protein isoforms—change in abundance between different conditions (Huang et al., 2015). Importantly, though, researchers ultimately care about the absolute abundance of RNA transcripts. In other words, is there a change in the number of RNA molecules in a cell when the conditions change? However, current techniques are limited to reporting relative abundances of RNA molecules: the proportion of fragments generated by a sequencer that contain a given sequence (Fernandes et al., 2014; Lovell et al., 2011; 2015; Quinn et al., 2018b). This means that RNA-Seq is inherently compositional data, where relative proportions are the only information available, yet those are being used to draw conclusions about the absolute abundance of features (Fernandes et al., 2014; Lovell et al., 2011; 2015; Quinn et al., 2018b) (see **Appendix 2.1**). Several studies have raised the alarm on ways in which interpretation of the results can be distorted if RNA-Seq data are not properly treated as compositional (Lovell et al., 2015).

The first statistical problem in an RNA-Seq analysis lies in determining the origin of the fragments generated. There are two classes of tools available to solve this problem: (1) tools that use traditional alignments to determine the exact genomic location (**tophat2**, **bwa**, **STAR**, **HISAT2**, etc.) (reviewed in (Risso et al., 2014)); there are other tools that take these traditional alignments and estimate exon-, transcript-, or gene-level expression levels (reviewed in (SEQC MAQC-III Consortium, 2014)); (2) tools that probabilistically estimate transcript sets that are

compatible with producing the corresponding fragments using pseudoalignment and quantify the levels of transcript expression (**kallisto**, **salmon**, **sailfish**) (SEQC MAQC-III Consortium, 2014).

The second statistical problem, the focus of this chapter, is to compare the differences in samples collected under different experimental conditions (e.g. comparing cancer cells with control cells; comparing wild-type cells with mutant cells). We will refer to this second step as "differential analysis." A number of tools are available for differential analyses (**DESeq2**, **edgeR**, **limma-voom**, etc.), using continuous or count data (reviewed in (Ferreira et al., 2014)). One recently developed tool, **sleuth**, utilizes the bootstraps produced by the quasi-mapping tools to estimate the technical variation introduced by the inferential procedure (SEQC MAQC-III Consortium, 2014).

It is a recognized need to normalize and transform the data before conducting differential analyses. Multiple strategies have been developed to meet this need, including quantile normalization (Jia et al., 2017; Pimentel et al., 2017), the trimmed mean of M-values (**TMM**) method used by **edgeR** (SEQC MAQC-III Consortium, 2014), the median ratio method used by **DESeq** and **DESeq2** (Risso et al., 2014), and the **voom** transformation used by **limma** (Piras and Selvarajoo, 2015). In addition, multiple units are used when modeling and reporting RNA-Seq results (Lovén et al., 2012), including counts (Fernandes et al., 2014; Lovell et al., 2011; 2015), CPM (Fernandes et al., 2014; Gloor et al., 2017), FPKM (Chen et al., 2015), and TPM (Gloor et al., 2017). Importantly, all of these strategies, even those that are focused just on the counts for each feature, utilize units that are really proportions, which belies the fact that RNA-Seq data are compositional (Ejigu et al., 2013; Rudnick et al., 2014) (see **Appendix 2.1**). Furthermore, all of these normalization strategies assume that the total RNA content does not change substantially across the samples (see (Martín Fernández et al., 2011) for a review). This assumption allows users

to leap from the inherently relative information contained in the dataset to the RNA copy number changes in the population under study without quantifying the actual RNA copy numbers. However, there are biological contexts where this assumption is not true (SEQC MAQC-III Consortium, 2014), and it is unclear how much change can occur before distorting results when the datasets are not considered as compositional during analyses.

If information is available about negative controls (e.g. spike-ins, validated reference genes), then such information could be used to anchor the data. This has been done in several studies, where the use of spike-ins led to a radically different interpretation of the data compared to the standard pipeline (Martín-Fernández et al., 2003). In one study, the RUVg approach was designed to use this reference information to normalize RNA-Seq data, as part of the RUVSeq R package (Martín-Fernández et al., 2003). There have been recommendations to include spike-ins as part of the standard protocol (Martín-Fernández et al., 2003). However, Risso et al. observed significant variation in the percentage of reads mapping to spike-ins, as well as discordant global behavior between spike-ins and genes (Martín-Fernández et al., 2003). While spike-ins are often used in single-cell RNA-Seq applications, they are not routinely used in bulk RNA-Seq experiments.

John Aitchison developed an approach to compositional data with the insight that ratios (or log-transformed ratios, called "log-ratios") capture the relative information contained in compositional data (Aitchison, 2008). There are three requirements for any analytical approach to compositional data: scale invariance, subcompositional coherence, and permutation invariance (see **Appendix 2.2**) (van den Boogaart and Tolosana-Delgado, 2013). It was recently demonstrated that correlation, a widely used measure of association in RNA-Seq analysis, is subcompositionally incoherent and may lead to meaningless results, and an alternative called "proportionality" was

proposed (Lovell et al., 2015). A tool was previously developed to apply compositional data analysis to differential analysis, called ALDEx2 (Fernandes et al., 2014). However, ALDEx2 is not well-suited for utilizing the bootstraps generated by the pseudoalignment tools and is unable to detect any differentially expressed features when there are less than five replicates (Quinn et al., 2018a). Therefore, it is necessary to develop a compositional approach for other tools.

Two tools are commonly used to simulate RNA-Seq: polyester and RSEM-sim (Frazee et al., 2015; Li and Dewey, 2011). These tools require the input of estimated counts per transcript and the expected fold changes between groups. However, without considering the data as compositional, protocols used to simulate RNA-Seq data result in the total read counts being confounded with the condition, such that one condition will have on average a greater depth compared to the other condition (for example, see Supplementary Table S2 of (Pimentel et al., 2017)). A protocol that simulates many changing features and at the same time yields similar sequencing depth per sample, is lacking, but could be done using principles from compositional data analysis. This challenge, along with the one above, both motivated the present work.

Here, we present **absSimSeq**, a protocol to simulate RNA-Seq data using concepts from compositional data analysis. This protocol allows us to directly model large global shifts in RNA content while still maintaining equivalent sequencing depths per sample. Further, we developed a normalization approach that uses negative control features (e.g. spike-ins) with log-ratios, which we call "compositional normalization". We created an extension of sleuth, called **sleuth-ALR**, to use compositional normalization, both to predict candidate reference genes and to normalize the data. We also adapted already available methods to implement compositional normalization for other differential analysis tools. We then used absSimSeq to benchmark performance of differential analysis tools in the setting of either a small or large change to the total RNA content.

Within each setting, we compared the current normalization approaches versus either compositional normalization or the RUVg approach with spike-ins.

When there was only a small change in total RNA content, all tested tools had similar performance on simulated data, whereas sleuth, sleuth-ALR and limma had the best performance on real data. However, when there was a large change in the RNA content, either up or down, all tools had much better performance if compositional normalization with spike-ins was used. When analyzing a well-characterized real dataset which had a large change in total RNA content per cell (Lovell et al., 2015; Marguerat et al., 2012), only compositional normalization with a validated reference gene was able to capture the overall decrease in RNA transcription. Surprisingly, RUVg had poor performance, and the choice of normalization approach had a greater impact on performance than the choice of tool. Furthermore, both of the concerns about spike-ins raised by Risso and colleagues (Risso et al., 2014) are actually the expected consequences of how compositional data behaves between samples, though they do raise concerns about the proper protocol for including spike-ins.

In summary, we provide **absSimSeq** as a resource to generate simulated RNA-Seq datasets that more accurately reflect the behavior of real datasets. This will help future development of RNA-Seq analysis when testing performance. Furthermore, our work suggests that using compositional normalization with spike-ins or validated reference genes is essential for differential analyses of RNA-Seq data. When such information is missing, it raises major concerns about the limitations of drawing conclusions from the inherently compositional data of RNA-Seq and other "omics" techniques.

3.2 Results

3.2.1 Simulation of RNA transcript copy numbers, normalization, and performance of different tools

To test how changes in the total RNA content can affect performance of differential analysis tools, we developed **absSimSeq** to simulate RNA-Seq data (**Figure 3.1**). Because the experimental step of generating a library from samples in a real RNA-Seq experiment generates a compositional dataset, this protocol directly simulates that step, producing simulated compositional data. Using this protocol, we carried out three simulation studies, each having five experiments. In each study, current normalization methods were compared to compositional normalization and **RUVg**. A set of highly expressed spike-ins was used as the set of negative control features for compositional normalization and RUVg (see methods for details).

In the first study ("small"), only a small fraction of features was changed (5% of all transcripts), and the total RNA content was similar between experimental groups (<2% change) (**Table A2.1** and **Table A2.2**). This study was intended to simulate an experiment that fulfills the assumption of the current normalization methods. Under these conditions, all tools tested performed similarly whether using their current normalization approaches or using compositional normalization (**Figure 3.2A**; **Figure S2.2**, panel A).

In the second set of studies ("down" and "up"), many features were differentially expressed (20% of all transcripts), resulting in a large change in the composition, with the total RNA content decreased by ~33% or increased ~2.8-fold, respectively (**Table A2.1** and **Table A2.2**). Under such conditions, compositional normalization led to greatly improved performance for all tools compared to these tools using their current normalization methods (**Figure 3.2B-C**; **Figure S2.2B-**

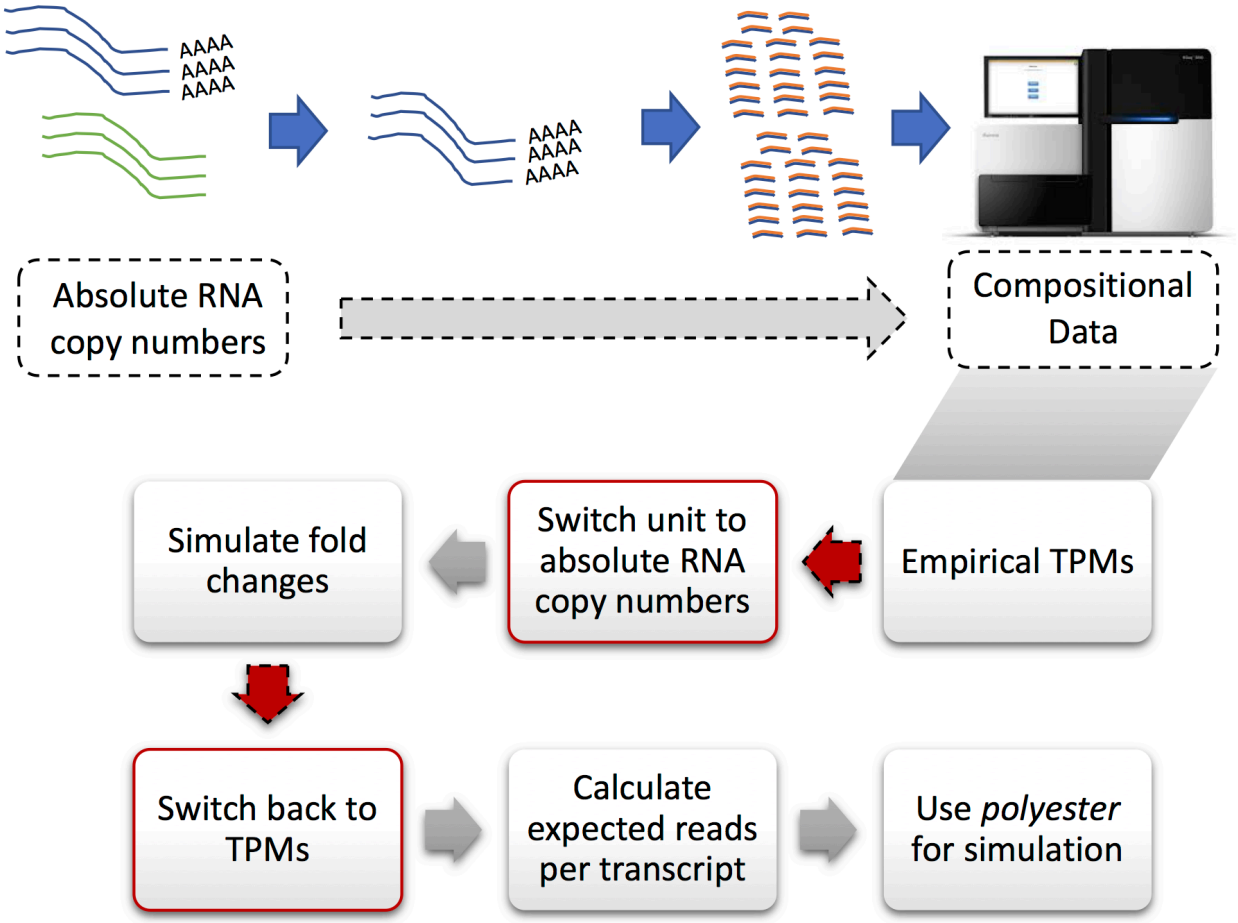

Figure 3.1: AbsSimSeq, A novel simulation protocol to model compositional RNA-Seq data*. All RNA-Seq experiments convert copy numbers per cell to relative abundances because of the selection step and because the depth of sequencing is arbitrary with respect to the total RNA present (top panel; see* ***Appendix 2.1****). The* ***absSimSeq*** *protocol simulates that conversion process (bottom diagram), with the key, novel steps highlighted in red. It takes the mean empirical relative abundances (in TPM units) from any dataset. It then makes a conceptual leap by assuming those values are copy numbers per cell. Then, the fold changes are simulated on these copy numbers for the experimental condition. After that, in the crucial step of the protocol, the copy numbers are re-normalized back to relative abundances to simulate what happens in the RNA-Seq experiment. From there, the expected reads per transcripts are calculated using relative abundances and the median of the estimated effective lengths from the original dataset. These are then submitted to the polyester R package for a negative binomial simulation.*

C). In contrast to current normalization methods and compositional normalization, the **RUVg** approach from **RUVSeq** resulted in the worst performance for edgeR and DESeq2 in all three studies (**Figure 3.2**), even though it used the same set of spike-ins as compositional normalization.

It is worth noting that, among the tools tested, sleuth and ALDEx2 performed the best when there were large compositional changes in the data (**Figure 3.2B-C**); ALDEx2 uses the IQLR transformation, which is a compositional approach designed to be robust in the presence of large changes to the composition. We also observed that for each transformation used by ALDEx2, it performed almost identically regardless of statistical test used (**Figure S2.3**). Finally, sleuth-ALR had similar performance whether TPMs or estimated counts were modeled, or if the Wald test or the likelihood ratio test was used (**Figure S2.4**).

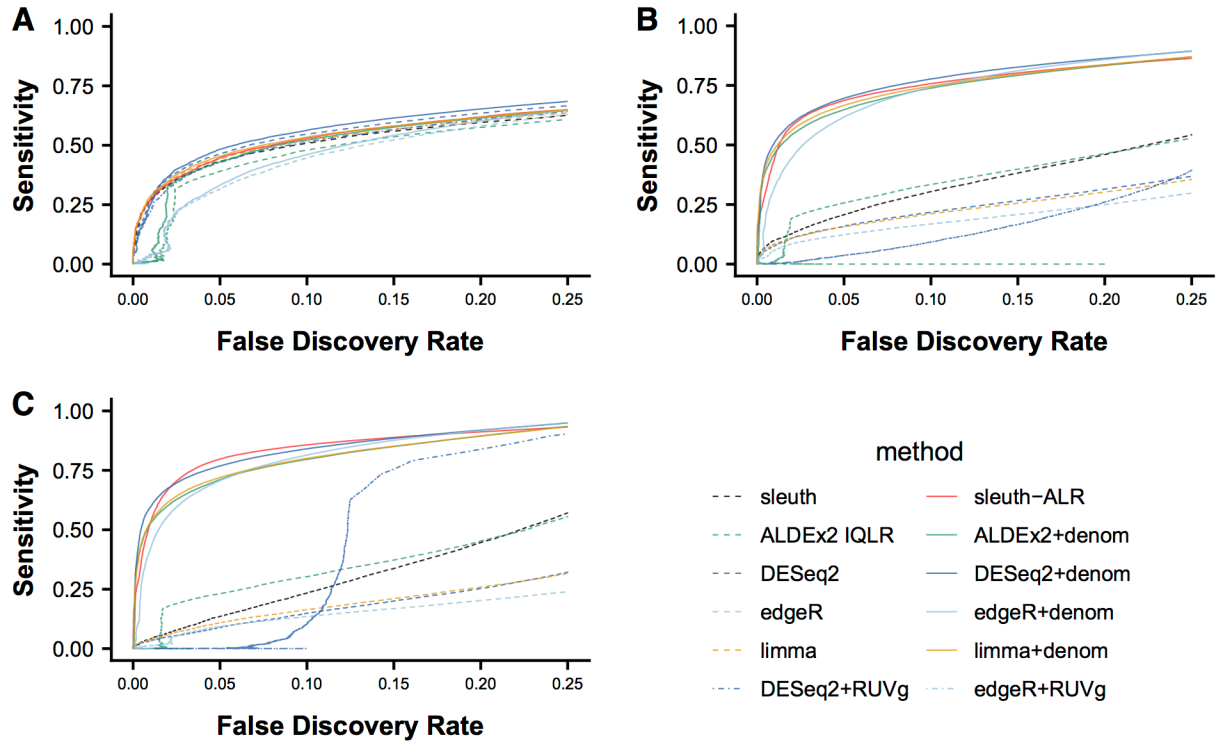

Figure 3.2: Compositional normalization markedly improves performance when there is a large compositional change. *The copy numbers were modeled using the estimated average abundances from the GEUVADIS Finnish women samples (N = 58). Each of the three studies consists of five simulations under specified global conditions: (A) the "small" group has the total average copy numbers per cell in each group roughly equal, (B) the "down" group has a large number of transcripts changing, with 90% of the transcripts down-regulated, and (C) the "up" group has the same number of transcripts changing as the "down" group, but 90% of the transcripts are up-regulated. The compositional normalization methods (solid lines) used a set of highly expressed spike-ins to illustrate. Average false discovery rate across the simulations within each group (n = 5) is shown on the x-axis, and average sensitivity is shown on the y-axis. The FDR range between 0 and 0.25 is shown. Note that edgeR+RUVg is not shown because it always had an FDR above 0.25. See* ***Figure S2.2*** *for the full range.*

3.2.2 Performance is not degraded by significant variation in individual spike-ins

A previous study reported that spike-ins had significant variation (Risso et al., 2014), raising a concern about their utility for normalization. In particular, they observed significant variation both between and within groups. In our simulated data, spike-ins were modeled similarly as other features, with over-dispersed variation, drawing from a negative binomial distribution. When estimating the percentage of transcript fragments mapping to spike-ins per sample, we also detected significant variation across samples (**Figure 3.3**). Importantly, in the "up" and "down"

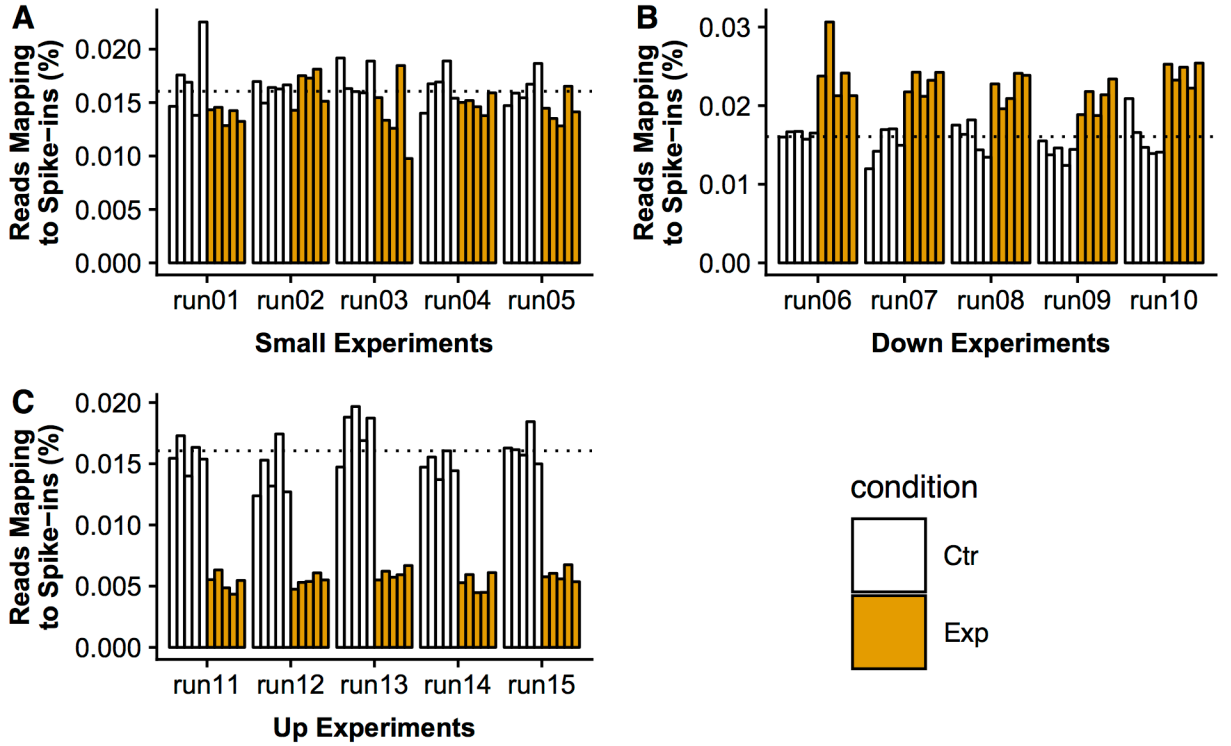

Figure 3.3: Spike-ins have significant within-group and between-group variation, despite improved performance when used for normalization. *Previous work expressed a concern about variation observed in spike-ins between samples. In each experiment, the 92 ERCC spike-ins from Mix 1 were simulated to have no change in copy numbers between the two conditions, as well as to have over-dispersed variation between samples, drawing from a negative binomial distribution. Plotted here is the percentage of all fragments that map to spike-ins, compared to the total number of fragments from the sample, in* ***(A)*** *the "small" study, with <2% change in the total RNA in each condition;* ***(B)*** *the "down" study, with a ~33% decrease in total RNA in the experimental condition; and* ***(C)*** *the "up" study, with a ~2.8-fold increase in total RNA in the experimental condition. The dotted line represents the expected percentage of fragments mapping to spike-ins in the control group. Across all experiments, there is significant within-group variation; in the "down" and "up" studies, there is also significant between-group variation. The latter is to be expected given the compositional nature of the data (see* ***Appendix 2.3****).*

studies, there were systematic differences between groups, similar to what was observed in the MAQC-III study and in the zebrafish study previously analyzed (Risso et al., 2014). These systematic differences between groups are expected given the compositional nature of the data (see **Appendix 2.3** and Discussion). We further observed a large spectrum of estimated fold changes across individual spike-ins, with a systematic asymmetry in the distribution of fold changes in the experiments from the "up" and "down" studies (**Figure A2.5**). Despite the significant variation of spike-ins, individually and collectively, spike-ins led to greatly improved performance when used for normalization (**Figure 3.2**). Consistent with previous work (Munro et al., 2014; SEQC MAQC-III Consortium, 2014), our results suggest that the ratio information contained in spike-ins are collectively robust to variation, and that spike-ins can be used for sample-wise normalization.

3.2.3 sleuth-ALR has best self-consistency and negative control performance among compositional normalization methods

To confirm that compositional normalization performs similarly to current methods in the context of real data, we repeated the analyses using data from the original study on sleuth (Pimentel et al., 2017). The first test was the "self-consistency" test using data from (Bottomly et al., 2011). We reasoned that a tool should provide consistent results from an experiment, whether a few samples per group are sequenced (in this case, n = 3 per group), or more samples per group are used (n = 7-8), as measured by the "true positive rate" (TPR) and "false discovery rate" (FDR). In this experiment, true positives were defined as hits identified in both the smaller and larger datasets, and false positives were defined as hits identified by the smaller dataset but not by the

larger dataset. Among all tools using compositional normalization, sleuth and sleuth-ALR with Wald test showed the best balance between the TPR and FDR (**Figure 3.4**; **Figure S2.6**). Limma-voom and sleuth/sleuth-ALR with the likelihood ratio test had the lowest FDR at the cost of a lower TPR. DESeq2 and edgeR both had higher FDR, and on average slightly lower TPR compared to sleuth-ALR. In contrast, the Welch and Wilcoxon statistics in ALDEx2 were unable to identify any hits in any of the "training" datasets, consistent with a recent benchmarking study

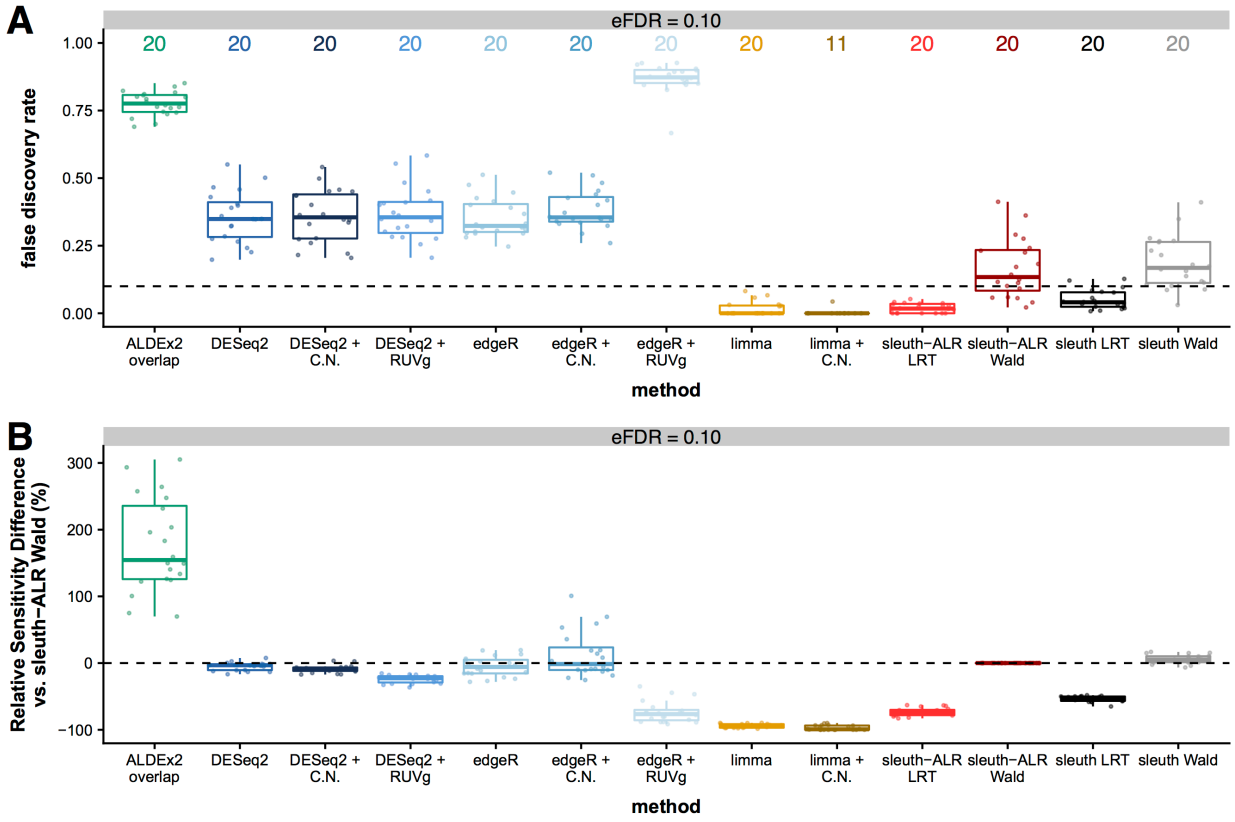

Figure 3.4: sleuth-ALR Wald has best balance of self-consistency between less and more data from same dataset. *Depicted is the Bottomly et al self-consistency test at the isoform level, with* ***(A)*** *the false discovery rate at three specified levels, and* ***(B)*** *the relative sensitivity as compared to sleuth-ALR with the Wald test. This extends the test from the original sleuth paper (Pimentel et al., 2017). A large dataset is split into a small "training" dataset (3 samples per group), and larger "validation" datasets. A "false discovery" in this test is defined as a hit identified in the "training" dataset but not in the larger "validation" dataset at the given FDR level, and a "true positive" in this test is a hit identified in both datasets at that FDR level. A tool performs well in this test if it can identify the same hits with less data, as well as control the "false discovery rate" at the specified FDR level. The full dataset was split twenty times. Note that the number above each tool in panel A is the number of "training" datasets out of twenty that identified at least one hit at the specified FDR level. See* ***Figure S2.6*** *for the results at the FDR levels of 0.01 and 0.05.*

(Quinn et al., 2018a) (data not shown), suggesting that they have greatly reduced power with less data (N = 3 samples per group). ALDEx2's "overlap" statistic was able to identify hits, but this led to the worst consistency (i.e. highest false discovery rate) among the tools tested. Finally, while DESeq2 with RUVg had similar performance to DESeq2 with compositional normalization, edgeR with RUVg had among the worst consistency.

Next, we tested the performance of compositional normalization on a negative control dataset, where there are no expected differentially expressed features. We repeated the null resampling experiment from the sleuth paper (Pimentel et al., 2017) using the GEUVADIS Finnish women dataset (n = 58) (The Geuvadis Consortium et al., 2013). Six samples were randomly selected (stratified by lab to minimize technical variation) and split into two groups, with the expectation of finding no hits. We found that sleuth-ALR with the likelihood ratio test performed similarly to sleuth and limma-voom (median number of false positives < 5) (**Figure 3.5**), and sleuth-ALR with the Wald test also showed good false positive control (median number of false positives = 10). In contrast, DESeq2 and edgeR with compositional normalization showed higher numbers of false positives (median of 71 and 66, respectively), and the "overlap" statistic for ALDEx2 showed the highest number of false positives (median of >5000 at the 0.1 FDR level).

3.2.4 Performance of compositional normalization on a dataset with a global decrease in transcription

To compare different tools in a real dataset with a large compositional change, we used the "yeast starvation dataset" (Marguerat et al., 2012). In this dataset, yeast cells were starved of a nitrogen source, inducing them to enter a reversible quiescent state without active cell division (Yanagida, 2009). Absolute copy numbers per cell were estimated for each mRNA by being

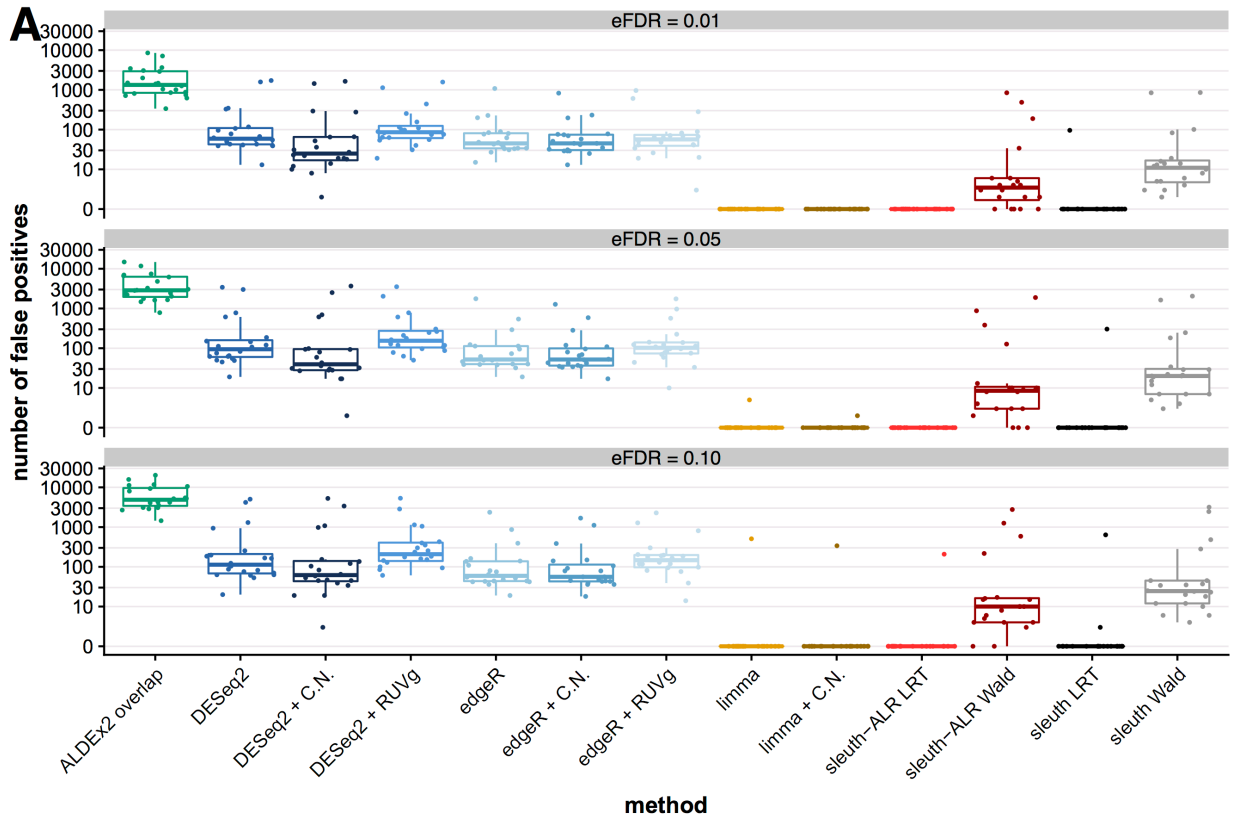

Figure 3.5: sleuth-ALR and limma perform best on the GEUVADIS null dataset. *Depicted is the null experiment at the isoform level* ***(A)****. This also extends the test from the original sleuth paper (Pimentel et al., 2017). The data were from the lymphoblastoid cells of 58 Finnish women, a relatively homogeneous population, taken from the GEUVADIS project (The Geuvadis Consortium et al., 2013). Data from six women were resampled from the larger dataset, stratifying by lab to minimize technical variation, and then randomly split into two groups to simulate a "null experiment". The number of false positives, defined as any hits, are reported here based on twenty rounds of resamplings. A tool performs well in this experiment by minimizing the number of hits reported. ALDEx2 used the IQLR transformation; all "C.N." methods and sleuth-ALR used compositional normalization; all "RUVg" methods used RUVg for normalization.*

normalized to a collection of 49 mRNAs that were quantified using NanoString nCounter (Kulkarni, 2001). Our re-analysis clearly shows a large global decrease in RNA content (**Figure 3.6A**), with ~95% of genes decreasing in copy numbers per cell in the starvation group versus control, confirming that the dataset has a large compositional change (**Figure 3.6B**). On the contrary, analyses using previously developed methods failed to identify this pattern of changes in gene expression, only reporting equivalent numbers of hits up- and down-regulated transcripts (**Table 3.1**).

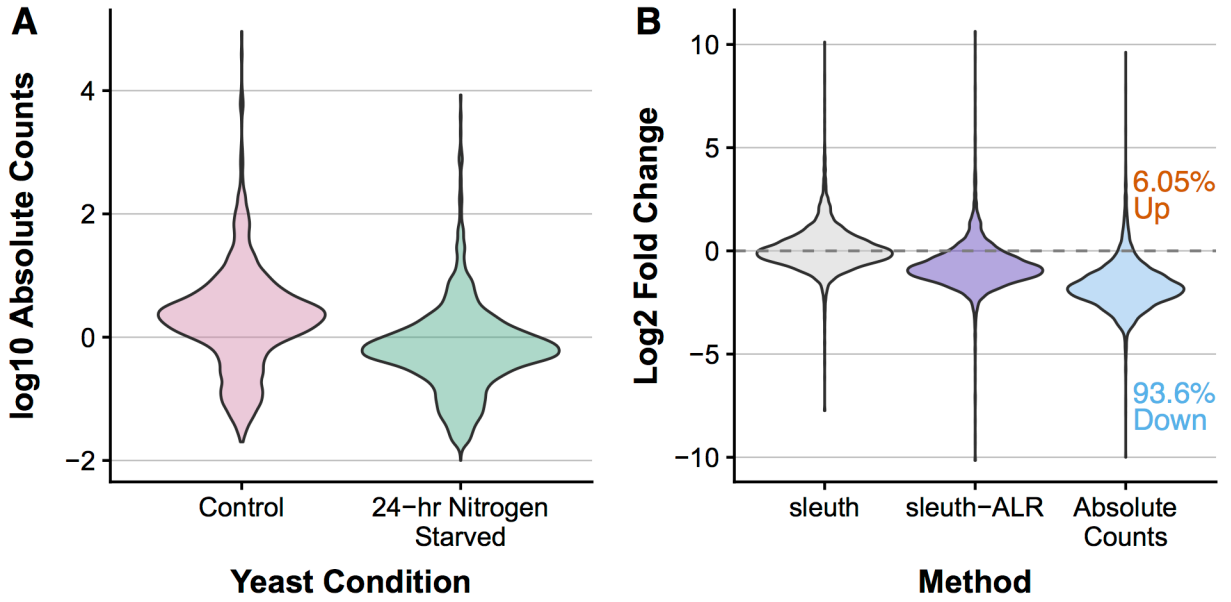

Figure 3.6: A yeast starvation study shows a large global decrease in RNAs. (A) *A violin plot showing the distribution of absolute counts in control yeast cells (pink) and yeast cells starved of their nitrogen source for 24 hours (green). The data were from Marguerat et al (Marguerat et al., 2012). The absolute counts were estimated by normalizing RNA-Seq data to a panel of reference genes whose copy numbers were quantified using the NanoString nCounter assay. As can be observed, there is a global decrease in the RNA present.* ***(B)*** *a comparison of log2 fold changes calculated using a standard RNA-Seq pipeline (the example shown here in gray is kallisto + sleuth), and the log2 fold changes calculated using compositional normalization (sleuth-ALR, shown in purple) or directly from the estimated absolute counts (in blue). As shown, the vast majority of genes were observed to be downregulated when estimating from the absolute counts. The standard RNA-Seq approach misses this global shift, but compositional normalization is able to identify it.*

Next, we examined the importance of using negative control features with compositional normalization. We first analyzed the data using the gene with the most constant proportion across all samples (as measured by coefficient of variation), rqc1 (Pombase: SPAC1142.01). In this context, compositional normalization missed the global pattern of down-regulation; instead, it reported a similar number of hits compared to other tools using current normalization methods, both up-regulated and down-regulated (**Table 3.1**). We then selected a gene with approximately constant expression (opt3, Pombase: SPCC1840.12) as the denominator for compositional normalization. Compared to current methods, compositional normalization using a validated reference gene was able to identify the global decrease in transcription observed by previous

analyses of the data (Lovell et al., 2015; Marguerat et al., 2012) (**Table 3.1**). All tools tested (ALDEx2, DESeq2, edgeR, limma-voom, and sleuth-ALR) were able to identify a similar number of hits when using compositional normalization. In contrast, while RUVg was also given the same validated reference gene, surprisingly it had greatly reduced power and was unable to capture the global pattern of down-regulation.

Tool	Up-Regulated Genes	Down-Regulated Genes
ALDEx2 ALR overlap	**719**	**4496**
ALDEx2 IQLR overlap	2716	2677
DESeq2	2424	2415
DESeq2 + C.N.	**632**	**4344**
DESeq2 + RUVg	**698**	**1001**
edgeR	2621	2536
edgeR + C.N.	**582**	**4290**
edgeR + RUVg	**109**	**90**
limma	2603	2527
limma + C.N.	**573**	**4332**
sleuth	2529	2554
sleuth-ALR (trend)	2692	2225
sleuth-ALR	**614**	**4356**
Absolute Counts	**522**	**5751**

Table 3.2: Only compositional normalization (C.N.) accurately reflects global decrease in the yeast starvation study. *This table shows the number of hits identified by each tool using default settings and kallisto-calculated estimated counts and abundances. "Sleuth-ALR trend" used rqc1 (Pombase: SPAC1142.01) as a denominator; this gene had the most consistent abundance (TPM value) across all samples. The compositional normalization methods (all tools in red: "ALDEx2 ALR overlap"; "sleuth-ALR"; all "+C.N." tools) used opt3 (Pombase: SPCC1840.12) as a denominator; this gene was considered a "validated reference gene". "RUVg" for edgeR and DESeq2 (in blue) also used opt3 as a negative control gene. Only compositional normalization methods, using opt3, were able to accurately reflect the severe global decrease observed in the data, as shown by the number of genes showing down-regulation of the absolute counts. Note that ALDEx2 Welch and Wilcoxon statistics yielded <5 significant hits.*

3.3 Materials and methods

3.3.1 absSimSeq approach to simulating RNA-Seq data

See **Figure 3.1** for a summary diagram of our protocol for **absSimSeq**. When generating RNA-Seq data, the key experimental step which requires a compositional approach is when the actual changes in the RNA content are sampled using an equal but arbitrary amount of RNA by the library preparation process, resulting in a dataset of proportions. To simulate this shift from count data to compositional data, the **absSimSeq** protocol starts with a set of transcripts and their TPMs, either defined by the user or estimated from real data. It then conceptually shifts from considering transcripts per million (a proportion) to considering copy numbers per cell, i.e. the number of transcripts present in each cell (the absolute count unit of interest). It then simulates the fold changes expected to occur between groups directly on the copy numbers, which may or may not result in a substantial change in the total RNA per cell. The next key step is then converting these new expected copy numbers back to TPMs to represent the expected proportion of each transcript that would be present in an equal aliquot taken from each group. These new TPMs are then converted to expected counts per transcript based on their lengths and the desired library sequencing depth, and those expected counts, along with user-defined or estimated parameters for variance within each group, are then submitted to the R package **polyester** to simulate an RNA-Seq experiment.

AbsSimSeq also has the option to add spike-ins to the simulated experiment. In our studies, the ERCC ExFold Spike-in mixes are used to define which sequences are included and in what proportions. The user can define what percentage of the transcripts should be coming from spike-ins and which mix to use.

3.3.2 Simulation of copy numbers for this study

To model a simulated dataset after real data, we took an approach modified from Patro et al (Patro et al., 2017) and Pimentel et al (Pimentel et al., 2017). To estimate the mean and variance for the control group for our simulation, we wished to use a population without expected biological changes within the group. We thus used as a proxy the largest homogeneous population in the GEUVADIS data set, a set of 58 lymphoblastoid cell lines taken from Finnish women. We estimated transcript abundances using kallisto and human Gencode v. 25 transcripts (Ensembl v. 87), and then estimated negative binomial parameters (the Cox-Reid dispersion parameter) using DESeq2. We next took the mean TPMs from this dataset, for input into **absSimSeq**.

Three simulation studies were performed, with five simulation experiments in each study. The "small" study was intended to simulate experiments where there was no substantial change in the copy numbers per cell per group. The "down" and "up" studies were designed to simulate experiments where there was a large compositional shift, with the total copy numbers either decreasing or increasing.

To simulate differential expression, we first applied a filter where the transcript had to have a TPM value of at least 1. We then randomly and independently assigned each filtered transcript as either not changing (i.e. fold-change of 1) or differentially expressed, using a Bernoulli trial with varying probability of success (5% of all transcripts for the "small" study; 20% for the "down" and "up" studies). For each differentially expressed transcript, a truncated normal distribution was used to simulate the fold change, with a mean of 2-fold, a standard deviation of 2, and a floor of 1.5-fold. A Bernoulli trial was then used to choose either up-regulation or down-regulation with varying probability of success (70% down for the "small" study, chosen to produce roughly equal total RNA in each group; 90% down for the "down" study; 90% up for the "up" study).

The estimated null distribution and the simulated fold changes thus defined the mean copy numbers per cell for the control group and the experimental group, respectively. These copy numbers were then converted back to TPM. Because TPM is proportional to the estimated counts divided by effective length (Pimentel et al., 2017), the TPMs were multiplied by the effective lengths and then normalized by the sum to get the expected proportion of reads per transcript per condition. This was then multiplied by a library size of 30 million reads to get the expected reads per transcript per condition. This matrix of expected reads and the Cox-Reid dispersion parameters estimated from the GEUVADIS dataset were used as input for the **polyester** package (Frazee et al., 2015) to simulate 5 samples in each group, with a random variation of about 2-3% introduced into the exact sequencing depth used. The dispersion parameters for the spike-ins was set to the median dispersion of all transcript that had a mean TPM within 5% of the TPM for the spike-in.

Table A2.1 summarizes the simulation parameters and the number of transcripts that are differentially expressed, and **Table A2.2** shows the average global copy numbers per cell per condition for each of the fifteen runs. Note that the experimental group in the "up" study had a ~2.8-fold increase, on average, in the total RNA copy numbers per cell. This is less than the 5.5-fold increase in total mRNA observed after over-expressing the oncogene c-Myc (See the normalized data in Table S2 in Lovén et al., 2012)). The experimental group in the "down" study had a ~33% decrease in the total RNA copy numbers per cell. This is less than the decrease in total RNA observed in the yeast dataset (See Supplementary Table S2 from (Marguerat et al., 2012)).

3.3.3 Implementing a compositional approach for differential analysis tools: the Log-ratio transformation

To allow tools to use negative control features, like spike-ins or validated reference genes, in a compositional manner, we present a method that uses what is called the "additive log-ratio" (ALR) transformation. This was proposed by John Aitchison to address problems analyzing compositional data. He demonstrated that any meaningful function of compositions must use ratios (Aitchison, 2008). Further, he proposed the use of log-ratios to avoid the statistical difficulty arising from using raw ratios. ALR is the simplest of the transformations proposed by Aitchison and others in the field of Compositional Data Analysis. In ALR, if there are D components in a composition, then the D-th component is used as the denominator for all of the other D-1 components.

Formally, if T is a set of D transcripts, then $\mathrm{x} = \{\mathrm{x_t}\}_{\mathrm{t}\subset\mathrm{T}}$ defines the relative abundance of the t-th transcript in the composition, with $\sum_{\mathrm{t}=1}^{\mathrm{D}} \mathrm{x}_t = \mathrm{C}$, where C is some arbitrary constant (e.g. 1 million for TPMs). These relative abundances are proportional to the units commonly used in RNA-Seq (RPKM, TPM, etc.) (Pachter, 2011). The ALR transformation takes a component to be used as the denominator, analogous to the "reference gene" used in qPCR experiments (see "How to interpret the results" below). This forms a new set of $\mathrm{D}-1$ transformed log-ratios,

$$\log_2 \frac{\mathrm{x_1}}{\mathrm{x_D}}, \log_2 \frac{\mathrm{x_2}}{\mathrm{x_D}}, \ldots, \log_2 \frac{\mathrm{x_{D-1}}}{\mathrm{x_D}}$$

that can then be used for downstream statistical analyses. If one wishes to use a collection of features (e.g. a panel of validated reference genes; a pool of spike-ins), then the geometric mean, g(x), of those multiple features can be used on all D components:

$$\log_2 \frac{x_1}{g(x)}, \log_2 \frac{x_2}{g(x)}, \ldots, \log_2 \frac{x_D}{g(x)}$$

If this is a subset of features, this is called the "multi-additive log-ratio" (malr) (Quinn et al., 2018a). If the geometric mean of all of the features are used, this is the "centered log-ratio" (CLR) transformation introduced by Aitchison and is used in the default mode of ALDEx2 (Fernandes et al., 2014). These are all options available for use in sleuth-ALR.

Log-ratio transformations are undefined if either the numerator or denominator is zero. **Sleuth-ALR** has implemented an imputation procedure for handling zeros that minimizes any distortions on the dependencies within the composition. See **Appendix 2.5** for more details.

3.3.4 How to choose a denominator for compositional normalization and how to interpret the results

The proposed interpretation of the results generated by **sleuth-ALR** is simple: whatever the denominator is, the results show how all the other features change relative to that feature or features. This is consistent with the permutation invariance requirement for compositional data analysis (see **Appendix 2.2**). For example, if GAPDH is selected as the reference feature, the results show how every other gene is changing relative to GAPDH.

What is important for interpretation of the results, though, is whether negative controls are available or not. Thus, if there are one or more features which are known *a priori* to be negative controls—a validated reference gene, a pool of spike-ins—these are natural choices for use as a

denominator in **sleuth-ALR**, either using a single feature, or using the geometric mean of multiple features. Since the copy numbers of these features are expected to be constant between samples, there is now an anchor for the relative proportions between samples.

If negative controls are unavailable, though, we propose identifying one or more features that have the most consistent proportion across all samples. For **sleuth-ALR**, we chose the coefficient of variation as the metric to measure consistency in proportion. Without external information, though, it is unknown if this "consistent feature" is indeed also consistent in copy numbers. If there is a global change in RNA content, this feature would represent the average change. All current normalization methods assume there is no such change, and so make corrections to the data to remove any perceived change; they are therefore mathematically equivalent to this proposed approach (see (Quinn et al., 2018b) for a full mathematical proof). However, this approach has the advantage of making explicit the implicit and necessary interpretation (how are features changing relative to the selected reference feature(s)?). It also provides a feature or set of features that can be used as a reference gene for follow-up validation.

3.3.5 How sleuth-ALR fits into the current sleuth pipeline

See **Figure S2.1** for the pipeline and how it compares to the current pipeline. In the current sleuth pipeline, estimated counts of transcripts from kallisto or salmon are first normalized by the DESeq2 median-of-ratios method (Anders and Huber, 2010), and then transformed on the natural log scale with a 0.5-fragment offset to prevent taking the logarithm of zero and to reduce the variability of low-abundance transcripts (Law et al., 2014; Pimentel et al., 2017). With the additive log-ratio transformation, the size factor is replaced by the estimated expression of the chosen denominator, and the offset is replaced by the imputation procedure. Once a denominator is

chosen, zero values are imputed, the ratios between each feature and the denominator is calculated, and the data is then transformed on a log scale, which can then be used directly in the sleuth model. Our implementation simply replaces the current normalization and transformation functions with the function provided by the **sleuth-ALR** package. For ease of interpretation, the modeling can be done on TPMs directly (using the *which_var* argument for *sleuth_fit*), though the previous choice of modeling the estimated counts can also be used.

3.3.6 Compositional approach for the other tools

ALDEx2 has an explicitly compositional approach, and it solves the imputation problem by simulating bootstraps on the data using the Dirichlet-multinomial distribution, which will never yield zero values. It then calculates estimated statistics by examining the differences between groups within each bootstrap. Its default option is to use the CLR transformation, but it has several options for other choices of denominator. A recent paper examined **ALDEx2**'s performance using these different options (Quinn et al., 2018a); and its results suggested that the IQLR ("interquartile range log-ratio transformation") provided the best balance of performance across real and simulated datasets, with respect to accuracy and computer time. This transformation uses the subset of all features that, after using the CLR transformation, have a sample-wise variance in the interquartile range. Theoretically, this transformation is robust to many features changing either up or down. This and the CLR transformation were used as the normalization methods tested in our study. It can also take a predefined subset of features and uses the geometric mean of those features within each sample; this was the approach taken when utilizing spike-ins for compositional normalization.

DESeq2 uses the median-of-ratios method (Anders and Huber, 2010). If one wishes to calculate a single size factor to normalize each sample, it is calculated by the *estimateSizeFactors* function. This function has a *controlGenes* option, which allows the user to define a set of features that are expected to have constant expression across all samples. A recent review demonstrated that the DESeq2 size factor is mathematically equivalent to the compositional log-ratio proposed by Aitchison (Quinn et al., 2018b). If a dataset has known negative control features (e.g. spike-ins), these can be used to calculate a DESeq2 size factor similar to what is calculated by **sleuth-ALR** or **ALDEx2**. For **DESeq2**, **edgeR**, and **limma**, we calculated DESeq2 size factors using the *estimateSizeFactors* function with designated negative control features (spike-ins for the simulated data; a validated reference gene for the yeast starvation dataset).

3.3.7 Pipeline to analyze simulations

The simulated data (FASTA files) were analyzed by **kallisto** for downstream use by all of the tools tested. Spike-ins from ERCC Spike-in Mix 1 were included for the simulations (2% of the total RNA), and so were used as the set of features known to have constant expression between samples. Previous studies observed that only highly expressed spike-ins had consistent ratios across samples (Munro et al., 2014; SEQC MAQC-III Consortium, 2014). Thus, we selected spike-ins that had an average log2 concentration of at least 3 between both mixes. This filter results in a set of 47 spike-ins that were used for compositional normalization and for **RUVg** from **RUVSeq**. **RUVg** was used with **DESeq2** and **edgeR** to test its ability to use spike-in information using its own approach.

Filtering is an important issue for managing the accuracy of estimation. Different pipelines make different decisions about what features to filter. To allow the tools to be compared fairly, the

same set of filtered transcripts were tested in all tools, defined by those transcripts that passed the standard sleuth filter of having at least 5 estimated counts in at least half of the samples. DESeq2's default functionality to use independent filtering and Cooks' outlier filtering did not significantly impact its performance on the simulated data (data not shown), so these were left on.

3.3.8 Experiments from the original sleuth paper

To see if compositional normalization would produce similar results with fewer replicates, we repeated the self-consistency experiment as described in the sleuth paper (Pimentel et al., 2017). Briefly, we used the Bottomly et al dataset (Bottomly et al., 2011), randomly split the 21 samples into a small training dataset consisting of 3 samples in each condition, and a large validation dataset consisting of remaining samples. The "truth" set of features was defined by the hits identified in the larger validation dataset. This was repeated 20 times. At each of three FDR levels (0.01, 0.05, 0.1), we compared the smaller dataset against the larger dataset, and plotted the estimated FDR and sensitivity relative to sleuth-ALR. Since spike-ins were not used in this experiment, and it is unknown if there was any significant change in the total RNA between the groups, the denominator for compositional normalization was chosen based on which feature had the lowest coefficient of variation across the whole dataset. Zfp106-201 (ENSMUST00000055241.12 in Ensembl v. 87) was used as the denominator for **sleuth-ALR** in all datasets. This was also used for RUVg. The IQLR transformation was used for ALDEx2, and otherwise the current normalization methods from the original sleuth paper were used.

To test the performance of compositional normalization when analyzing a negative control dataset, we also repeated the null resampling experiment as described in the sleuth paper (Pimentel et al., 2017). Briefly, we used the Finnish samples from the GEUVADIS dataset, and randomly

subsampled the data into twenty null experiments with 3 samples in two groups. This subsampling was stratified by lab to minimize technical variability that may have occurred between labs. Because of the homogeneous population and minimized technical variation, the expectation is that there would be zero differentially expressed features. The null experiments were analyzed, and the number of false positives was plotted at the transcript-level and gene-level. The same denominator was used for compositional normalization in **sleuth-ALR** across all twenty of the null experiments: SRSF4-201 (ENST00000373795.6) for the transcript-level and SRSF4 (ENSG00000116350) for the gene-level. This transcript and gene were determined to have the respective lowest coefficient of variation across all of the samples used for this experiment.

3.3.9 Pipeline to analyze yeast dataset

To test the different tools and normalization approaches on a real dataset, we chose a well-characterized "yeast starvation" dataset (Marguerat et al., 2012). In this dataset, yeast cells were cultured in two conditions: (1) freely proliferating using Edinburgh Minimal Medium; (2) the same medium without a nitrogen source (NH_4Cl), resulting in the cells reversibly arresting into a quiescent state. Two samples from each condition were processed for poly-A selected RNA-Seq or for total RNA (no selection or depletion step). A collection of 49 mRNAs were selected for absolute quantification using the NanoString nCounter, which uses a fluorescent tagging protocol to digitally count mRNA molecules without the need for RNA purification. The results were normalized to external RNA controls to estimate copy numbers per cell of each mRNA. We used the absolute counts summarized in Supplementary Table S2 of (Marguerat et al., 2012) as a basis for selecting opt3 (Pombase: SPCC1840.12.1) as the gene with the smallest coefficient of variation for estimated absolute counts among all samples. This can be considered a validated reference

gene. Thus, it was used with methods utilizing negative control features (**sleuth-ALR**, **RUVg**, and other tools using **DESeq2**'s estimateSizeFactors with controlGenes argument) to normalize the data. **Sleuth-ALR** was also tested using rqc1 (Pombase: SPAC1142.01.1), which was selected as having the smallest coefficient of variation for raw abundances (TPM values) across all the samples. This gene represents the "average global trend" or "average global change" in the data, as discussed in "How to choose a denominator" section above.

To re-analyze the RNA-Seq data, we downloaded the *Schizosaccharomyces pombe* genome cDNA FASTA file from ftp.ensemblgenomes.org (Fungi release 37). This was used as the reference for generating the kallisto index. Each tool was then run using default settings.

3.3.10 Availability of data and code

The yeast starvation dataset was taken from Marguerat et al (Marguerat et al., 2012) from ArrayExpress at accession E-MTAB-1154, and the absolute counts were taken from Supplementary Table S2 from (Marguerat et al., 2012). The GEUVADIS Finnish data can be found at ArrayExpress using accession E-GEUV-1, using the samples with the population code "FIN" and sex "female". The Bottomly et al data (Bottomly et al., 2011) can be found on the Sequence Read Archive (SRA) using the accession SRP004777. Human annotations were taken from Gencode v. 25 and Ensembl v. 87, mouse annotations were taken from Gencode v. M12 and Ensembl v. 87, and yeast annotations were taken from Ensembl Genomes Fungi release 37. The code and vignette for **absSimSeq** can be found on GitHub at www.github.com/warrenmcg/absSimSeq, the code and vignette for using sleuth-ALR can be found at www.github.com/warrenmcg/sleuth-ALR, and the full code to reproduce the analyses in this chapter can be found at www.github.com/warrenmcg/sleuthALR_paper_analysis. Here are the

versions of each of the software used: **kallisto** v. 0.44.0, **limma** v. 3.34.9, **edgeR** v. 3.20.9, **RUVSeq** 1.12.0, and **DESeq2** 1.18.1; the version of **polyester** used is a forked branch that modified version 1.14.1 with significant speed improvements (found here: www.github.com/warrenmcg/polyester); the version of **sleuth** used is a forked branch that modified version 0.29.0 with speed improvements and modifications to allow for **sleuth-ALR** (found here: www.github.com/warrenmcg/sleuth/tree/ speedy_fit); the version of **ALDEx2** used is a forked branch that modified version 1.10.0 to make some speed improvements and to fix a bug that prevented getting effects if the ALR transformation with one feature was used (found here: www.github.com/warrenmcg/ALDEx2). All R code was run using R version 3.4.4, and the full pipeline was run using snakemake.

Preface to Chapter 4

Haipeng Cheng needs to be acknowledged for generating the HEK FUS KO cells. Jianwen Deng needs to be acknowledged for the initial EM observations that prompted this study, as well as the individual mitochondria study that appears in Figure 4.5, as well as some of the qPCR work that appears in this study (Figure 4.8).

CHAPTER 4: Molecular and Bioinformatics Studies of the Role of FUS in Mitochondrial Function

4.1 Introduction

Several studies have examined FUS targets using high-throughput methods in a variety of biological contexts. Early work examined FUS targets using microarrays (Blechingberg et al., 2012; Camats et al., 2008; Colombrita et al., 2015; Fujioka et al., 2013; Ishigaki et al., 2012) and Affymetrix ChIP-Chip or RIP-Chip (Colombrita et al., 2012; Tan et al., 2012), whereas later studies employed RNA-Seq (Kapeli et al., 2016; Lagier-Tourenne et al., 2012; Masuda et al., 2015; Nakaya et al., 2013; Reber et al., 2016; Scekic-Zahirovic et al., 2016; Schwartz et al., 2012; van Blitterswijk et al., 2013) and RASL-Seq (Scekic-Zahirovic et al., 2016; Sun et al., 2015) for expression and alternative splicing profiling, and PAR-CLIP (Hoell et al., 2011), HITS-CLIP(Ishigaki et al., 2012), iCLIP (Rogelj et al., 2012), CLIP-Seq (Kapeli et al., 2016; Lagier-Tourenne et al., 2012; Masuda et al., 2015) and RIP-Seq (Reber et al., 2016) for identification of FUS targets. As mentioned in section 1.3, the challenge with identifying FUS targets is that FUS does not seem to have a clear sequence preference or structural motif for binding sites. Instead, it appears to have affinity for RNAs dependent on length, with RNA seeding oligomers of FUS (Schwartz et al., 2013; Wang et al., 2015b). Further, work using CAP-Seq, GRO-Seq, and Poly-A-Seq have examined the role of FUS in selection of transcription start sites and poly-adenylation sites (Masuda et al., 2015; 2016b).

All told, there have been 57 datasets generated from 34 published and unpublished studies on FUS using these high throughput technologies. Many of these papers used an overrepresentation analysis to examine what processes were enriched among statistically

significant hits, with a majority identifying synaptic maintenance and function as major processes affected. However, besides the mention of mitochondrial function in two of the studies (both in HEK293 cells) (Hoell et al., 2011; Schwartz et al., 2012), there has been no comment on a possible link between FUS and mitochondrial function in any of these high-throughput studies. We hypothesized two reasons for this: (A) it has only recently been noted that pathway enrichment analyses are sensitive to the choice of background genes to include (Timmons et al., 2015); and (B) it is possible that FUS has a strong global effect on RNA processing, such that there is a compositional shift that would distort differential expression estimates (see Chapter 3).

Before this work, we had preliminary evidence to suggest that FUS had an endogenous role in mitochondrial function. In HEK cells, mitochondrial pathways were identified as significantly enriched with genes affected by FUS knockdown (Schwartz et al., 2012). In our previous work with FUS overexpression models and in post-mortem brain samples, we observed a pool of FUS that endogenously localized to mitochondria in HEK cells and in the brains of healthy humans (Deng et al., 2015). We then observed smaller mitochondria in the brains of FUS KO mice using electron microscopy; this size decrease was similar to what was observed in our overexpression models (unpublished observations).

We thus reasoned that, because of the established roles of FUS in RNA regulation and RNA localization (see section 1.3), FUS may have a role in regulating mitochondrial-associated transcripts. Further, because there is evidence to suggest that loss-of-function mechanisms contribute to the pathogenesis of FTLD and ALS, studying the endogenous role of FUS in mitochondrial function may help contribute to our understanding of the diseases. We thus sought to use a knockout model of FUS to study more carefully its role in mitochondrial regulation.

To this end, we generated HEK FUS KO cells using CRISPR/Cas9. Using electron microscopy, we observed a similar size decrease in mitochondria after FUS KO (unpublished observation). Our aims in this study were (A) to characterize the mitochondrial dysfunction in HEK FUS KO cells; (B) to identify from previous high-throughput studies putative FUS-regulated mitochondrial-associated genes; and (C) see if these mitochondrial-associated genes regulated the observed mitochondrial dysfunction.

4.2 Results

4.2.1 Using sleuth-ALR results in a dramatic re-interpretation of the global expression changes in FUS RNA-Seq Studies

Of the many RNA-Seq datasets available for FUS, we decided to focus on three: a FUS KO dataset generated from the embryonic brains of a brand new mouse model (not the Hicks or Kuroda model) (Scekic-Zahirovic et al., 2016), a FUS knockdown dataset generated in neuronal-like N2a mouse cells (Masuda et al., 2015), and a mutant FUS (R521G) dataset generated from patient-derived iPS cells differentiated into motor neurons (compared to cells derived from wild-type siblings) (Kapeli et al., 2016). They represent a broad range of neuronal model systems across two species, and they also represent both loss-of-function and gain-of-function models, including one directly relevant to disease.

We analyze each dataset using the same pipeline that culminated in a traditional normalization approach (the standard normalization provided with **sleuth** (Pimentel et al., 2017)), and a compositional normalization approach (using **sleuth-ALR** discussed in Chapter 3 (McGee et al., 2019)). For all three datasets, we chose the same reference gene as the denominator for sleuth-ALR: GAPDH. The study in N2a cells (Masuda et al., 2015) explicitly mentioned (A) that there are no FUS binding sites in GAPDH, and (B) GAPDH expression was

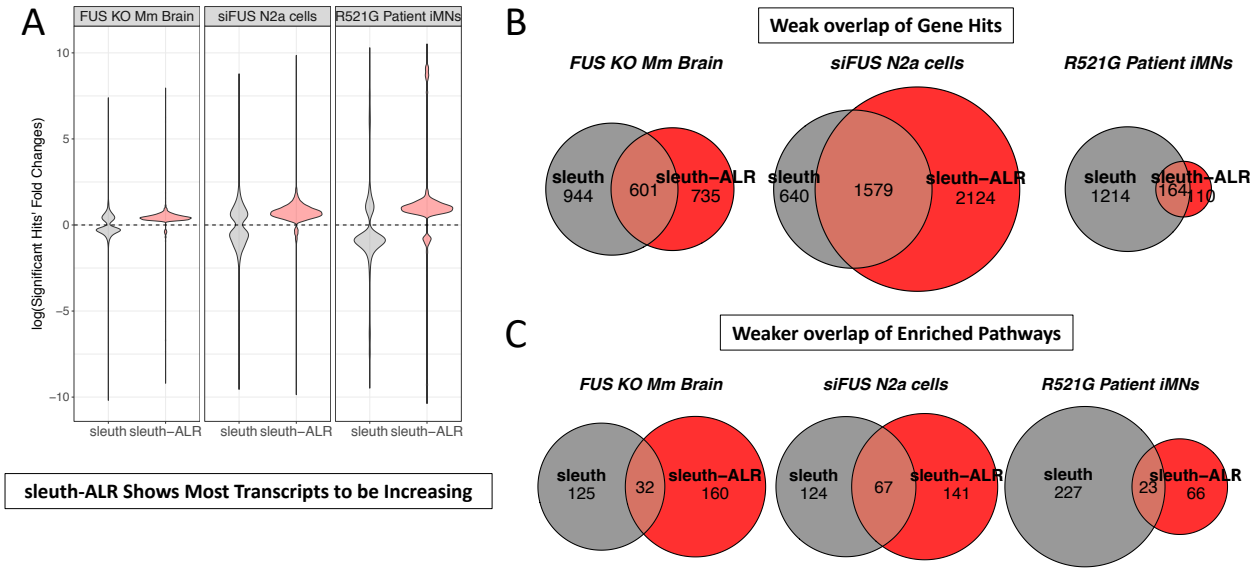

Figure 4.1: Dramatic Reinterpretation of FUS RNA-Seq Datasets after sleuth-ALR analysis. (A) *A violin plot of the distribution of fold changes of transcripts identified as differentially expressed by sleuth (gray) or sleuth-ALR using GAPDH as a reference gene (red). Sleuth-ALR results show an extreme skew toward up-regulation.* ***(B)*** *A Venn diagram showing the modest overlap of differentially expressed transcripts identified by the original sleuth analysis versus the sleuth-ALR GAPDH analysis* ***(C)*** *A Venn diagram of pathways identified as enriched with differentially expressed genes (p-value aggregation) using topGO. There is an even weaker overlap here between the two analyses.*

unchanged after FUS knockdown (though they did not show this data). Even though GAPDH has not been validated, it is commonly used as a reference gene, including in the iPSC-derived motorneurons study (Kapeli et al., 2016). We wanted to see if there were any significant differences in the interpretation of the data after manipulation of FUS.

Surprisingly, in all three datasets, normalizing the RNA-Seq data to GAPDH expression resulted in a dramatic re-interpretation of the global expression patterns (**Figure 4.1**). In particular, whereas the standard analysis identified a symmetric pattern of up- and down-regulated transcripts, the GAPDH-normalized data showed an extremely skewed distribution of differentially expressed transcripts, with almost all transcripts showing up-regulation (**Figure 4.1A**). This pattern was observed in all three datasets. There was only a modest overlap when comparing the list of hits from each analysis (**Figure 4.1B**), and there was even less overlap

when examining which pathways were identified as affected by FUS manipulation (**Figure 4.1C**). If GAPDH is a true reference gene, this global pattern of regulation indicates that FUS globally acts as a repressor of mRNA expression, whether at the transcriptional or post-transcriptional level.

4.2.2 HEK FUS KO cells have no change in proliferation or galactose sensitivity

A previous study observed that cell viability (as measured by the MTT) was compromised in the mouse cell line NSC-34 and in HEK cells after shRNA FUS knockdown (Ward et al., 2014). Given that observation our observation of the smaller mitochondria in HEK FUS KO cells, we hypothesized that FUS KO would compromise mitochondrial bioenergetics. However, when we measured cell number and apoptotic cells using an automated cell counter, we saw no difference between HEK FUS KO cells and control cells (**Figure 4.2**).

If there was no difference in cell proliferation under unstressed conditions, we then sought to determine if there would be a difference if the cells were forced to just rely on mitochondrial respiration. The galactose sensitivity assay measures the ability of cells to proliferate when they are forced to use galactose, a sugar that can only be metabolized using mitochondrial respiration. Treating the cells with antimycin A and galactose served as a negative control. We again observed no difference between FUS KO cells and control cells (data not shown). In summary, no change in proliferation and no change in

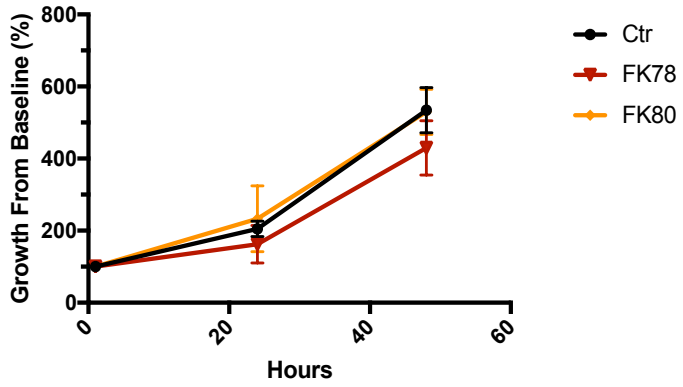

***Figure 4.2: No change in proliferation after FUS KO**. Cells were grown for 48 hours and cell counts were estimated using an automated cell counter.*

galactose sensitivity makes it unlikely that the mitochondria are severely compromised under these conditions.

4.2.3 HEK FUS KO cells have no change in mitochondrial respiration or glycolysis

If mitochondria are not severely compromised in HEK FUS KO cells, we next sought to determine if there is a more subtle defect in bioenergetics. To do this, we used the Seahorse assay, which can assess in a comprehensive way the bioenergetics profile of cells (Brand and Nicholls, 2011). The assay can directly measure oxygen consumption (as a proxy for mitochondrial function) and extracellular acidification (as a proxy for glycolysis). In our hands, we saw no differences across the whole profile for both oxygen consumption and glycolysis (**Figure 4.3**).

4.2.4 HEK FUS KO cells have no change in mitochondrial membrane potential, but increase in mitochondrial mass

If smaller mitochondria have no change in bioenergetics, it is possible that they have a compromise in maintaining mitochondrial membrane potential (MMP). We thus tested whether mitochondrial membrane potential was changed in FUS KO cells, using FACS and the TMRE

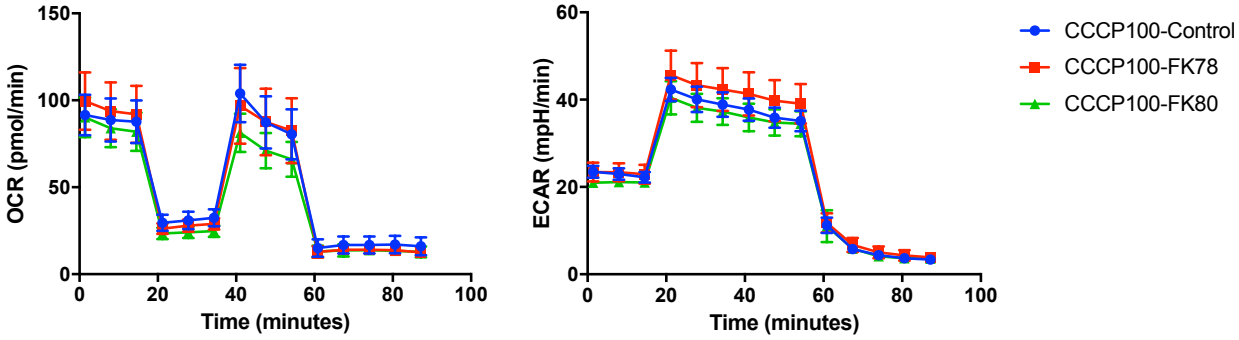

Figure 4.3: No change in bioenergetics after FUS KO. *The Seahorse assay was used to measure oxygen consumption (OCR, left panel) and extracellular acidification (ECAR, right panel). See methods for drugs and concentrations used. No differences were observed across the whole profile.*

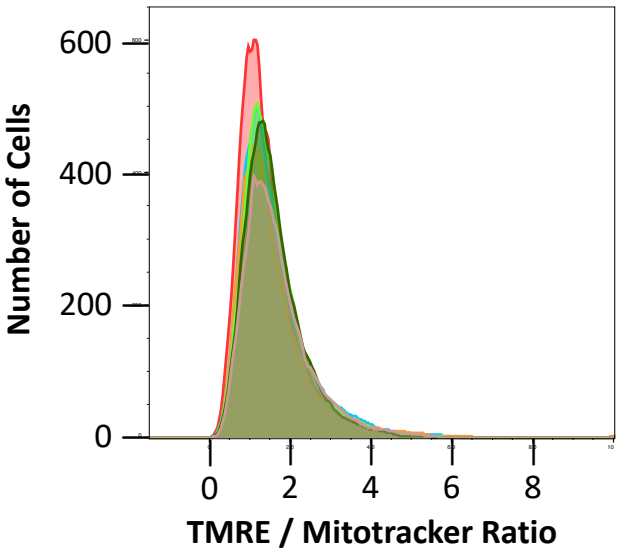

Figure 4.4: No change in MMP after FUS KO. *Shown are the distribution of individual events detected from triplicate samples of FUS KO cells and control cells. What was measured was the normalized TMRE/MitoTracker ratio. As shown, no differences were observed between conditions.*

MMP-sensitive dye (Cottet Rousselle et al., 2011). Because this dye is possibly sensitive to mitochondrial size and density, we normalized the TMRE signal to MitoTracker Green FM, which is known to be insensitive to MMP (Cottet Rousselle et al., 2011). In our hands we did not see a difference in normalized TMRE signal between FUS KO cells and control cells (**Figure 4.4**). We did, however, see an increased in MitoTracker Green signal, suggesting that there was increased mitochondrial mass in HEK FUS KO cells (**Figure 4.5**).

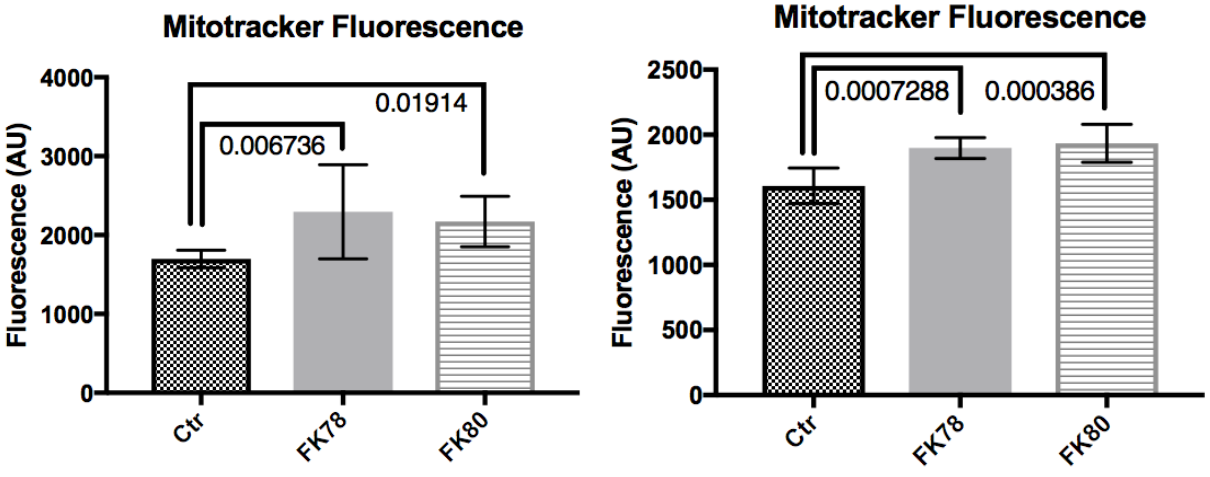

Figure 4.5: FUS KO cells have increased mitochondrial mass*. HEK cells were stained with MitoTracker Green FM. Fluorescence intensity is a proxy for mitochondrial mass. Both FUS KO clones have increased mitochondrial mass. Shown are two independent experiments. The posthoc p-values are displayed with each comparison.*

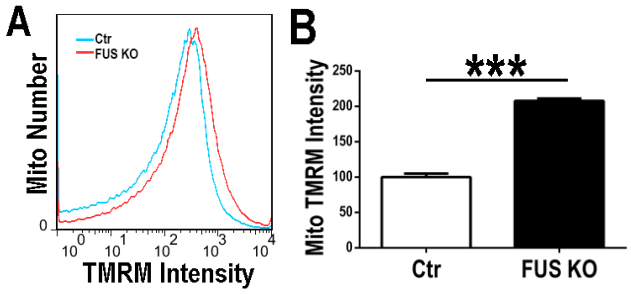

Figure 4.6: Individual mitochondria from FUS KO cells have increased MMP. *Purified mitochondria were stained with TMRM and measured using FACS. (A) A histogram of TMRM intensities for control cells and FUS KO clone #78. (B) Quantification of the mean intensity average across three experiments*

4.2.5 Isolated mitochondria from FUS KO cells have increased mitochondrial membrane potential

The increase in mitochondrial mass despite a smaller mitochondrial size suggested that there might be a compensatory effect occurring, where individual mitochondria might be compromised, but the cells respond by producing more. To test this, we isolated mitochondria from FUS KO and control HEK cells, and repeated the MMP experiment, looking at the signal from individual mitochondria. Surprisingly, we observed an increased in TMRM signal in FUS KO cells versus control HEK cells, suggesting a hyper-polarization and an improved function of mitochondria (**Figure 4.6**).

4.2.6 qPCR validation failed when using standard approach

A previous study had identified mitochondrial function as the most strongly affected pathway in HEK cells after FUS knockdown (Schwartz et al., 2012). We thus sought to determine which mitochondrial-associated genes would be consistently decreased in HEK cells and the nervous system. The original data for that study was not available, so we used the RNA-Seq data from the neuronal models as a proxy. After re-analyzing the RNA-Seq datasets from FUS neuronal models using the standard approach, we identified several mitochondrial-associated genes that were consistently decreased across multiple studies. We thus sought to determine if these genes were also decreased in HEK cells. However, out of 10 genes tested, we only detected changes in two genes (**Figure 4.7**). One was a decrease in MTHFD2, which the

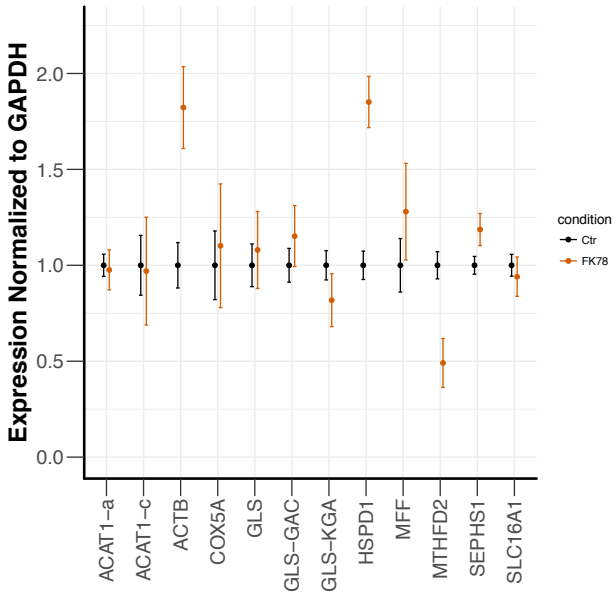

Figure 4.7: qPCR validation fails with hits identified by standard analysis*. Shown are eleven transcript-specific or gene-level primer sets to validate expression changes identified in the neuronal datasets. Actin B was also included as an additional reference gene, but its expression was different from GAPDH. Of the eleven primer sets tested, only one (MTHFD2) was validated with a similar direction of change*

neuronal RNA-Seq data had predicted to be decreasing. The other was an increase in HSPD1/HSP60, which we had already identified as a significant interactor with FUS in a previous study (Deng et al., 2015). However, the neuronal datasets had predicted a *decrease* in HSP60, suggesting inconsistent regulation of HSP60 in neuronal versus non-neuronal cells.

4.2.7 qPCR validation succeeded with sleuth-ALR

After developing the sleuth-ALR approach and re-analyzing the previous datasets using GAPDH as a denominator, all of the genes were previously tested were no longer considered significant, and a new set of genes were identified as significantly changing across multiple neuronal studies. We therefore tested this new set of genes in our HEK cells, using GAPDH as a reference gene. We selected 12 genes, all of which were predicted to increase after FUS KO from the neuronal datasets. Out of the 12 genes tested, eight of them showed an increase in expression relative to GAPDH (**Figure 4.8**). Importantly, all of the genes detected as changing had a consistent direction of change as compared to the neuronal datasets.

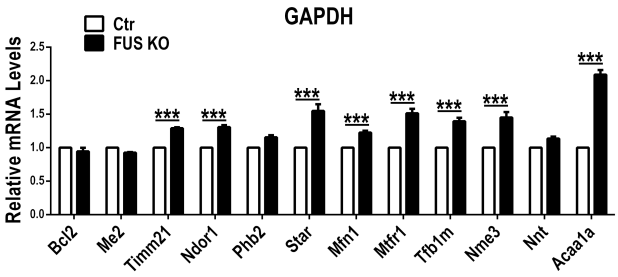

Figure 4.8: Hits identified by sleuth-ALR had much higher rate of validation in FUS KO HEK cells. *Shown is a comparison of control cells with FUS KO clone 78. Of an additional twelve genes tested, eight of them were validated.*

4.2.8 GAPDH may also be increased in HEK FUS KO cells

To validate GAPDH as a reference gene in our HEK FUS KO cells, we designed a modified qPCR protocol that uses an exogenous RNA "spike-in" control (see Methods, section 4.3.10). This was inspired by previous work showing the efficacy of using spike-in controls to normalize data in situations where the global content of RNA significantly changes (Lovén et al., 2012). These spike-in RNAs have similar properties to other polyadenylated do not have significant homology with any known sequences in the human genome. We isolated RNA from a pre-defined number of cells, and then we added the spike-in RNAs in a pre-determined amount proportional to the number of cells. We also prepared a control sample that had the same amount of spike-in RNA but did not go through the RNA isolation. By taking equal *volumes* rather than equal RNA mass to perform reverse transcription, we could compare each sample to the control sample to test (A) whether there was any inhibition in the reverse transcription or qPCR reactions, and (B) to normalize each sample for any variation in total RNA content or efficiency from the RNA isolation protocol. The spike-in RNAs can then serve as a proper reference to assay any validate any reference genes.

Using this modified protocol, we observed in a preliminary experiment a two-fold increase in GAPDH expression between HEK FUS KO cells versus control cells (**Figure 4.9**). These results need to be reproduced in additional experiments, but it suggests that GAPDH may

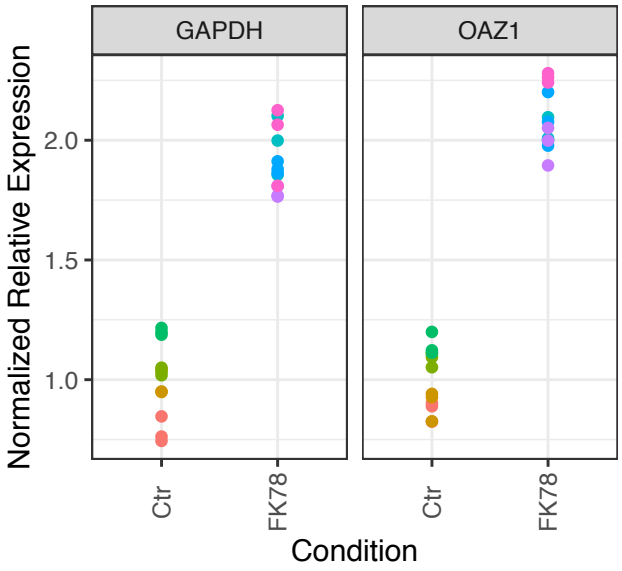

Figure 4.9: GAPDH may be up-regulated in HEK FUS KO. *A modified protocol was used to include exogenous spike-in RNA proportional to the number of cells used for RNA isolation, and then equal volumes were used for reverse transcription. The spike-in is then used as a validated reference feature to validate other reference features. Relative to the spike-in, both GAPDH and another commonly used reference gene OAZ1 both show a 2-fold up-regulation.*

be increased in our FUS KO cells. If this is true, then mitochondrial-associated genes that failed validation in Figure 4.7 may also be increasing.

4.3 Methods

4.3.1 Identification of eligible FUS and RNA-Seq datasets

To identify eligible datasets for inclusion in a re-analysis of FUS, a focused search was performed on Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/), ArrayExpress (https://www.ebi.ac.uk/arrayexpress/), and the DNA Data Bank of Japan (DDBJ; https://www.ddbj.nig.ac.jp/index-e.html). The following search terms were used: 'FUS OR TLS OR "Fused in Sarcoma" OR "Translocated in Liposarcoma"'. The resulting hits were manually reviewed to identify all relevant datasets. Datasets related to studies of FUS fusion proteins were excluded. All other datasets, regardless of method, were included. This search was performed in December of 2016.

For additional datasets, a focused literature search was performed using the following search terms on Pubmed: '(FUS OR TLS OR "Fused in Sarcoma" OR "Translocated in

Liposarcoma") AND ("RNA-Seq" OR "RNA-Sequencing" OR "RNA sequencing" OR "Microarray")'.

4.3.2 Pipeline for re-analysis of FUS RNA-Seq datasets

Studies that used RNA-Sequencing were selected for re-analysis. Raw FASTQ files were downloaded from SRA or ERA using the study accession. Fragments were aligned without trimming to the appropriate transcriptome downloaded from Gencode (v25 for humans; vM13 for mice) or Ensembl (version 87; December 2016 release) using kallisto (Bray et al., 2016). For read lengths 75-nt or shorter, a k-mer length of 21 was used to build the kallisto index; otherwise, the default k-mer length of 31 was used. The shorter k-mer was necessary because smaller read lengths are less likely to map correctly if the default length is used (for example, if an error or variant occurs at position 25 in a read of length 50, there is no chance for the read to map with a k-mer of 31).

After kallisto quantification, standard sleuth (Pimentel et al., 2017) or sleuth-ALR (McGee et al., 2019) was used to perform transcript-level differential expression analysis. Gene-level differential expression was also done using p-value aggregation (Yi et al., 2018). If a reference gene was used in the study for follow-up validation, the most abundant transcript for that gene was used as the denominator for sleuth-ALR. Otherwise, GAPDH was used. GAPDH was chosen because (A) it was a commonly used reference across multiple studies, and (B) one study had stated that GAPDH was not a target of FUS nor was its expression level changed after FUS knockdown (Masuda et al., 2015).

After sleuth-ALR differential analysis, the topGO R package (Alexa et al., 2006) was used to calculate gene ontology enrichment of significant genes using its default algorithm. This method de-correlates the gene ontology terms with each other, which greatly reduces redundancy

in the list of enriched pathways. Analysis was focused on the "Biological Process" set of gene ontology terms, and gene sets with less than 5 genes were excluded. The list of significant genes was defined as all genes that had an FDR≤0.05 after p-value aggregation. The list of background genes were all genes that had at least one transcript that passed the standard sleuth filter (at least 5 counts in at least 47% of the samples).

A custom R script was then used to compare the results from standard sleuth analysis and sleuth-ALR analysis.

4.3.3 generation and culture of the HEK FUS KO model system

All HEK293T cells were cultured in DMEM (4.5g/L of glucose; Sigma) supplemented with 10% (vol/vol) fetal bovine serum (FBS), 1% (vol/vol) of Penicillin, Streptomycin (P/S), and Glutamate (Gibco) (unless otherwise noted) at 5% CO2 and 37°C. Cells were split at confluency using the standard protocol with trypsin (TrypLE Express, Gibco). These cells were treated for mycoplasma contamination using Plasmocin treatment (InvivoGen) following the manufacturer's instructions; the cells were confirmed to be mycoplasma free using Universal Mycoplasma Detection Kit (ATCC).

To generate FUS KO cells, we followed a published protocol (Ran et al., 2013). HEK cells were plated in 6-well plates containing medium without antibiotics at a density of 3×10^5/mL with 1 mL of media per well, and the cells were allowed to attach for 24 hours. They were then transfected with the pSpCas9(BB)-2A-puro plasmid and sgRNA targeting exon 1 of FUS (sequence: 5'- TGCGCGGACATGGCCTCAAA-3') using lipofectamine2000. The sgRNA was designed to minimize off-target effects. After 18 hours, the media was changed. After another 6 hours, the cells were trypsinized and re-plated in 10-cm dishes at a dilution of 1:1200. After 48 hours, puromycin (final concentration: 2.5 ug/mL) was added as a selection marker, and

then media with puromycin was changed every 1-2 days. After 1 week, clones were selected, and tested for FUS KO status using western blot. Two clones (78 and 80) were selected for downstream experiments. In experiments with one clone, the clone with the more extreme phenotype (78) was used.

4.3.4 proliferation assay

HEK FUS KO cells and control cells were grown to confluency in a 10-cm dish, and then trypsinized for 3 minutes until the cells were detached, quenched with pre-warmed media, spun at 300g for 5 minutes at 4˚C. Supernatant was removed, and the pellet was resuspended in 2 mL of pre-warmed media. Cell density was measured using a Cellometer (Nexcelom) following the manufacturer's protocol. Cells were then plated into 12-well plates in quadruplicate for each time point at a density of 1×10^5/well (1 mL media), using DMEM (4.5 g/L glucose) and 10% FBS and 1% P/S without glutamate or pyruvate. Cell densities were measured at 24 hours, 48 hours, and 72 hours using the Cellometer. Data presented is representative of three experiments.

4.3.6 galactose sensitivity assay

HEK FUS KO cells and control cells were seeded into 12-well plates as described above (section 4.3.4). This time, they were cultured with media containing DMEM and 10% and 1% P/S without pyruvate or glutamate, and either glucose (4.5 g/L) or galactose (4.5 g/L), and either treated with antimycin A (final concentration: 2 uM) or left untreated. Cell densities were measured at 24 hours, 48 hours, and 72 hours using the Cellometer.

4.3.7 Seahorse assay

The XFe96 Seahorse cartridge was hydrated using the provided Seahorse XF Calibrant overnight in a 37˚C humidified incubator using room air (low CO2). CellTak was applied to the Seahorse plate following manufacturer's instructions. Then, the HEK FUS KO cells and controls

were seeded at a density of 5×10^4 cells/well, spun in a centrifuge to get the cells to attach. They were then incubated in a non-CO2 incubator for at least 30 minutes before running the plate. Port 1 applied oligomycin at a final concentration of 2 uM, port 2 applied CCCP at a final concentration of 100 nM, and port 3 applied a combination of rotenone (9 uM), antimycin A (9 um), and 2-deoxyglucose (2 mM). All biological replicates were seeded onto the plate in at least quadruplicate to measure technical variation.

4.3.8 Mitochondrial membrane potential assay

The day before the experiment, HEK FUS KO cells or control cells were trypsinized and seeded into 12-well plates at a density of 1×10^5 cells/well (1 mL of media). On the day of the experiment, media was removed, and media containing the dyes were added: a mix of TMRE (final concentration: 25 nM) and MitoTracker Green FM (final concentration: 50 nM); at these concentrations, these dyes are known to be non-quenching (Mitra and Lippincott Schwartz, 2010; Perry et al., 2011; Schaefer et al., 2016). The cells were incubated in the dark for 30 minutes at 37˚C. The media was then removed, and the cells were tryspinized and collected into FACS tubes. These were spun in a centrifuge at 300g for 5 minutes at 4˚C. The supernatant was removed, and the pellet was resuspended in 300 uL of "staining buffer" (PBS plus 2% FBS). The cells were then transported on ice and in the dark to the Flow Cytometry core facility, and then measured on a BD LSRFortessa Analyzer. Gates were set up to filter out doubles and debris as well as to filter out non-specific signal from unstained conditions. At least 10,000 events were analyzed. Single-stain controls were used to confirm that no compensation was necessary for these dyes on this instrument. We also treated the cells with 10 uM of CCCP for 5 minutes as a depolarization control. The data was analyzed using FlowJo, by taking the ratio of the TMRE

signal and the MitoTracker signal for each event and subtracting the average signal from the CCCP-treated condition.

4.3.9 Measurement of membrane potential of purified mitochondria

For each condition, the mitochondrial fraction was isolated from two 10-cm plates using a previously published protocol (Deng et al., 2015). This purified fraction was resuspended in 1 mL of mitochondria assay (120 mM KCl, 5 mM KH2PO4, 1 mM EDTA, 1 mM MgCl2, 3 mM HEPES, pH 7.8, 1% BSA) and divided into 100 uL portions that were supplemented with 10 mM succinate. After 1 hour of incubation at 37°C, the mitochondria were stained for an additional 15 minutes with 1 mM MitoTracker Green to differentiate intact mitochondria from cellular debris and 100 nM TMRM to determine mitochondrial membrane potential. Samples were analyzed on a BD FACS Calibur machine with FlowJo software. At least 100,000 positive events were collected for each sample. Data represents three independent experiments.

4.3.10 qPCR

HEK FUS KO cells and control cells were grown to near confluency, and then total RNA was collected using Trizol, following the manufacturer's instructions. Isolated RNA was resuspended in 40 uL DEPC-treated water. RNA purity and concentration were measured using a NanoDrop 2000 Instrument (ThermoFisher). All samples had 260/280 ratios > 1.9 and 260/230 ratios > 2.0. Reverse transcription was carried out using the SuperScriptIII reverse polymerase, following manufacture instructions and a 1:1 mix of oligo-dT(18) primers and N6 random primers. For each batch of samples, the same input mass of RNA was used (1-5 ug). The cDNA was diluted to a final concentration of 5 ug/uL, and 10 ug was used as input for each qPCR reaction.

qPCR was either performed on a BioRad CXF96 instrument using the Power SYBR™ Green PCR Master Mix (Applied Biosystems), or on a Stratagene MX3005P instrument (Agilent) using the UltraSYBR (Low ROX) Master Mix (CWBio CW2601M). In both cases, a fast 2-step protocol was used with an initial melting step at 95˚C for 10 mins, then 40 cycles of 15 seconds at 95˚C followed by 30 seconds (BioRad) or 1 minute (Stratagene) at 60˚C. A melting curve was then done to check for product specificity, consisting of a dissociation step at 95˚C for 15 sec followed by 1 min at 55˚C and continual readings up to 95˚C with a 0.1˚C/second ramp.

4.3.11 modified qPCR

In this modified protocol, the same procedure is done with the following modifications. We first counted cells using a hemocytometer and isolated RNA from 1×10^6 cells per sample. The cells were spun in a centrifuge at 300g for 5 minutes at 4˚C; the supernatant was removed and 600 uL of Trizol was added to each sample. After this, 2×10^7 or 8×10^7 copies of Universal RNA Spike-in 1 or 2 (TATAA Biocenter), were added to the sample to serve as an external spike-in. The standard Trizol RNA extraction protocol was followed. The isolated RNA was resuspended in 40 uL of DEPC-treated water. RNA concentration and purity were measured on a NanoDrop 2000 instrument. Extraction and inhibition controls were made by adding the same amount of TATAA RNA Spike-in 1 or 2 (2×10^7 or 8×10^7 copies) to DEPC-treated water.

Then, instead of equal mass of RNA input for each sample, equal 4-uL aliquots (10% of total RNA extracted) were used for the reverse transcription step using the SuperScriptIII reverse polymerase, with a mix of 1:1 oligo-dT(18) and N6 random primers used. Included in each reaction (including the control samples) was 2×10^7 copies of Luciferase Control RNA (Promega)

to serve as an inhibition control. The cDNA was then diluted 1:10 before use with the instrument, taking 2 uL aliquots for each qPCR reaction.

4.3.12 Data analysis

All data generated from the Seahorse, Proliferation, and Galactose sensitivity assays were analyzed using a one-way ANOVA with Tukey post-hoc tests in GraphPrism 7. All qPCR data was processed using the LinRegPCR normalization method (Ramakers et al., 2003; Ruijter et al., 2009)(either the stand-alone software or the "slope" method in the *sliwin* function in the qpcR R package (Ritz and Spiess, 2008)). The LinRegPCR method estimates an amplification efficiency for each sample; the mean efficiency per gene was used for subsequent analysis. The processed data was then processed using R 3.4 and a custom R script, using a nested ANOVA model (technical samples nested in biological replicates) on the relative abundances calculated by ΔΔCt method (Pfaffl, 2001).

CHAPTER 5: Discussion and Summary

5.1 FUS, TDP-43, and miRNA biogenesis: what's next?

5.1.1 The limitations of our network approach

We set out to design a pipeline that would predict the processes that TDP-43 would regulate via its regulated miRNAs and their predicted target genes. There are some limitations to the approach taken: (1) if a target interaction was missing from even one sample, it was excluded from the set of targets analyzed for a miRNA; this made the analysis easier to interpret, but it excluded miRNA-mRNA interactions that are likely important for a subset of lung cancer samples; (2) only the most differentially expressed genes from the Fatiscan results were included for the FatiGO analysis; thus, any subtle but biologically important signals in the data were ignored; (3) we did not examine whether there were miRNA-mRNA interactions that were positively correlated, which can indicate miRNAs either enhancing mRNA expression (Orang et al., 2014) or acting in a "tuning" or noise-buffering capacity (Bartel, 2009; Noorbakhsh et al., 2013; Osella et al., 2011); (4) we cannot exclude the possibility of other miRNAs or other genes being the true cause of the changes we observe in the target gene. However, despite these limitations, this pipeline provides a clear set of hypotheses for future work to validate.

The resultant predicted causal interaction network provides a complex picture of the predicted impact of TDP-43 on the pathogenesis of lung cancer. One aspect that complicates analysis has to do with the opposing roles of alternative isoforms. Several of the genes have only one transcript predicted to be a target, and this leads to a context-specific effect on cancer pathogenesis. For example, our pipeline predicted miR-423-3p to have four important targets in LUSC. Prima facie, the results seem mixed because you have two genes that inhibit cell

migration (LCP2 (Baker et al., 2009) and ADRB2 (Yu et al., 2007)), one with mixed results with respect to cell migration (ITGA9 (Mostovich et al., 2011)), and one that is an oncogene that promotes cell migration (CRK (Sriram and Birge, 2011)). However, when one looks more closely at the transcript level, the CRK transcript that is targeted is CrkIII, the shortest isoform with a predicted structure that has a truncated SH3 domain (Sriram and Birge, 2011). Thus, it is reasonable to suppose that, given the established role of miR-423-3p in promoting cell migration, CrkIII acts as a competitive inhibitor of Crk signaling and that inhibition by miR-423-3p leads to restoration of Crk signaling that promotes cell migration. Future work would need to be done to confirm this hypothesis and tease out the other complexities of transcript-specific miRNA targeting.

Another complicating aspect was how the miRNAs that were hits in our pipeline have mixed roles in tumorigenesis. Of the 28 miRNAs reported as hits in either LUSC or LUAD samples, 22 have previous literature exploring their roles in cancer (12 miRNAs are indicated as suppressors and 10 are indicated as oncomiRs; see supplemental tables 7 and 8); there was a trend toward tumor suppressor TDP-43-regulated miRNAs being down-regulated (hypergeometric test p-value for LUAD (4/4 vs 40/85) and LUSC (7/9 vs 40/85) respectively: 0.045 and 0.054), which suggested that TDP-43 is a tumor promoter. However, this trend was not seen with oncomiRs overrepresented among up-regulated hits (p-values for LUAD and LUSC were 0.614 and 0.177 respectively).

5.1.2 Future possibilities

FUS was identified to regulate synaptic processes via FUS-regulated miRNAs, which is consistent with previous work (Ling, 2018). One important way to improve on this analysis is to apply the compositional normalization approach to this network analysis. It’s not clear if there is

a large global change in RNA content in our dataset that would result in distorting the observed fold changes. Given how dramatically the observe global pattern changed when using sleuth-ALR, more work needs to be done to carefully validate reference genes and determine the global pattern of FUS and TDP-43 regulation. The set of mRNA targets that are differentially expressed. An additional problem with studying tissue is that there is an additional layer of compositional data generated because of the various cell types present in the tissue. One could imagine a scenario where no differential expression occurred within a cell type, but that number of cells varied between conditions. In this situation, the change in cell composition results in observed differential expression independent of cellular changes. There are a few techniques available to try to resolve this both computationally (Mancarci et al., 2017; Xu et al., 2013), and experimentally (Miller et al., 2014; Wang et al., 2019b), so these could be implemented here.

Another intriguing angle to pursue is the recent line of work suggesting a connection between cancer risk and neurodegeneration risk discussed in the introduction (Murmann et al., 2018; Umansky, 2018). Since both FUS and TDP-43 are both intimately involved with miRNA biogenesis, it is tempting to hypothesize that they are important contributors to suppressing toxic small RNAs. TDP-43 is known to suppress transposable elements (Li et al., 2012), and a recent preprint provides evidence that FUS suppresses the production of snoRNA-derived small RNAs (Plewka et al., 2019). This would be an interesting line of inquiry to pursue in the future.

Finally, more work could be done to explore the “tuning” interactions of miRNAs with their targets. As mentioned above, in these interactions, a positive correlation would be expected, and these are completely unexplored. Recent papers have published methods of assaying paired mRNA and small-RNA expression profiles in single cells (Wang et al., 2019a; Xiao et al., 2018). Since correlations are not compositionally coherent (and thus are meaningless without a negative

control feature), this work would require including external spike-ins. This could prove to be a powerful method to measure the genome-wide network of miRNA-mRNA interactions.

5.2 Compositional Normalization: What's Next?

Compositional normalization performed similarly to current normalization methods when there was only a small change to total RNA (**Figure 3.2A; Figure 3.4; Figure 3.5**). This similarity is expected because, in this scenario, it is valid to assume that most features are not changing. However, when there was a large change in global RNA, compositional normalization using negative control features (spike-ins) had much better performance compared to the current normalization methods, for both simulated data (**Figure 3.2B-C**) and real data (**Table 3.1**). In this case, the assumption held by the current methods is violated, and this is likely what greatly reduced their performance. Further, although the IQLR transformation in ALDEx2 was designed to be robust to large changes in global RNA, it only modestly improved performance. This indicates that at least some of the features it selected and assumed to be unchanging were indeed changing, in both the simulated data and the real data.

The worse performance of current normalization methods is likely related to how fold changes behave in the absolute case versus the relative case. Current normalization methods assume that most features are not changing; one could equivalently assume that the total RNA per cell is unchanged (Evans et al., 2018). Thus, if the total RNA content changes, current methods will anchor the data on whatever this global change is. This results in a shift in the observed distribution of fold changes (see Supplementary Figure S16 of (Lovell et al., 2015)). Extreme changes will still be observed to have the same direction, but there will be a group of features that are changing less dramatically than the global change that will be observed to have the wrong sign, and many unchanged features will appear to be changing.

Interestingly, the choice of normalization method had a much greater impact on performance than the choice of differential analysis tool, validating a finding from an older study (Bullard et al., 2010). This was true both for the simulated data (**Figure 3.2B-C**) and for the yeast dataset (**Table 3.1**). In both cases, compositional normalization clearly outperformed current methods when analyzing a dataset with substantial changes to the total RNA content, whether simulated or real. However, this was only true when negative control features were used (**Table 3.1**). This indicates that the choice of tool is much less important than the choice of normalization and the availability of negative control features (spike-ins, validated reference gene) to properly anchor the data.

Surprisingly, RUVg had poor performance in both the simulated data and the experimental yeast dataset. It is unclear why this occurred, other than the likely possibility that it treats the data as count data rather than compositional data. More work would need to be done to see if RUVg could be modified to more accurately capture the global trend in the data from negative control features.

In our simulations, the total RNA content was either decreased by 33% or tripled, respectively. The overall change in the "up" study was less extreme than what was observed after c-Myc overexpression (Lin et al., 2012; Lovén et al., 2012). In that context, the researchers found a general transcriptional activation that was not captured by the traditional analysis of the RNA-Seq data, and required cell number normalization using spike-ins to see the overall trend of increasing gene expression; the total RNA content increase observed by RNA-Seq was ~5.5-fold (see Table S2 of (Lovén et al., 2012)). The overall change in the "down" group was less extreme than that observed after the Marguerat et al dataset, which observed an 88% decrease total RNA content when using normalized RNA-Seq data. How often large shifts occur in real datasets is

unclear because of how infrequently spike-ins or validated reference genes are used when generating data. Future work should determine more carefully how drastic composition changes need to be before performance starts to degrade for methods which assume that most features are not changing.

5.2.1 Results from Bottomly et al. self-consistency test and GEUVADIS null experiment

Sleuth-ALR had the best self-consistency (**Figure 3.5**), and sleuth-ALR and limma had the best performance in the negative control dataset (**Figure 3.6**). ALDEx2 was unable to identify any hits using three samples per group with the standard statistical methods (Wilcoxon and Welch), and its "overlap" statistic showed a very high FDR, indicating that its results were inconsistent between the "training" and larger "validation" datasets. This indicates that ALDEx2 may not perform well when there are few replicates per group. While this manuscript was in preparation, a recent benchmarking study came to the same conclusion (Quinn et al., 2018a). This behavior is likely due to the fact that the algorithm of ALDEx2 does not include any shrinkage of the variance. Variance shrinkage has been demonstrated to improve performance when there are few replicates (Law et al., 2014; Love et al., 2014; Wu et al., 2013). Interestingly, though, all three statistics used by ALDEx2 had similar performance on the simulated data (**Figure S2.3**), and the "overlap" statistic identified a similar set of hits in the real dataset as other compositional normalization methods (**Table 3.1**), suggesting that the "overlap" statistic may have utility in small datasets despite poor self-consistency or poor control of false positives in a negative control dataset. Future work could explore how to improve ALDEx2 performance for smaller datasets.

5.2.2 The lack of real datasets with verified global changes

It was difficult to identify an example of a real dataset, with clear-cut evidence of substantial changes to the total RNA content, that was also amenable to re-analysis using our pipeline. We were unable to re-analyze the previous data measuring the impact of c-Myc overexpression (Lovén et al., 2012) because the RNA-Seq dataset did not have technical or biological replicates. We were also unable to re-analyze the selective growth assay used in the ALDEx2 paper (Fernandes et al., 2014) because the raw data, which is necessary for our pipeline, was not publicly available. Other datasets have used spike-ins, but they had no other confirmatory data on the absolute copy numbers to confirm if the spike-ins accurately captured the global trend or not. This dearth of bulk RNA-Seq datasets with verified global changes speaks to how much the problem of neglecting to treat bulk RNA-Seq as compositional data has gone unrecognized in the community.

5.2.3 How to choose a denominator for compositional normalization and interpret the results

When using compositional normalization, regardless of which denominator is chosen, the interpretation of differential expression and fold-changes is "the change of feature X with respect to the denominator". Although all transformations are permutation invariant and therefore any chosen denominator will produce mathematically equivalent results (Aitchison, 2008), the choice of denominator has important implications for the interpretation of the results and for the downstream validation experiments.

If an experimenter has information about absolute copy numbers per cell in their experiment, they can readily use that information with compositional normalization. For example, if spike-ins are included proportional to the number of cells, as recommended in the c-Myc study (Lovén et al., 2012), those spike-ins can be used as the denominator. If one or more

reference genes are validated, as was done with the Yeast Starvation study (Marguerat et al., 2012), then a reference gene known to be approximately constant under the experimental conditions can be used. In principle, if qPCR is used to validate differential analysis results in this scenario, a predicted reference gene after using spike-ins or the validated reference gene used for compositional normalization would be the best choice for a reference gene.

What about experiments that do not have spike-ins or validated reference genes? Spike-ins have only slowly been adopted as a part of *Seq protocols (Hardwick et al., 2017). It has further been extensively documented that reference genes are frequently not properly validated (Bustin et al., 2009), and that expression of commonly used reference genes could change dramatically under certain circumstances (Barber, 2005; Rubie et al., 2005). There have been several techniques to identify reference genes using RNA-Seq data (Bin Zhuo et al., 2016; Van L T Hoang et al., 2017; Yim et al., 2015). Importantly, these techniques all find a feature that has an approximately constant proportion throughout all of the samples. However, researchers are usually attempting to identify a reference gene with approximately constant absolute copy numbers per cell throughout. In order to draw this conclusion, the techniques must make the same assumption that standard RNA-Seq analysis tools make, i.e. that the global RNA content remains constant in all samples, or that only a few features are differentially expressed. If many features are changes, features identified by these tools will only reflect the global change (up or down), rather than being approximately constant in absolute copy numbers per cell.

None of the compositional normalization methods solve this problem (for an example, see sleuth-ALR with the “trend” feature compared to the other methods in **Table 3.1**), because no tool can *in principle* solve this problem without access to external information. As described in a recent review article (Quinn et al., 2018b), no approach can formally recapitulate the

absolute data, and only approaches that are using truly constant features can adequately anchor the data to accurately estimate the true changes in the absolute data. In most datasets without spike-ins or validated reference genes, it is unknown if there is a significant change in the total RNA per cell. Thus, all that is left is how one feature behaves relative to another feature.

When one feature is used, there is a clear advantage to compositional normalization versus current methods because there is a clear interpretation of the results (i.e. how features are behaving relative to this feature), and because there would be a clear choice of reference gene for any qPCR validation downstream (for example, spp1 would be used in the yeast starvation study). Any other choice for qPCR reference gene would likely yield discordant results. Importantly, though, identifying a feature with approximately constant proportion, in the absence of information about the overall changes in RNA content, can still help experimenters identify important biology. This is analogous to the approach taken by Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). Its “competitive" null hypothesis leads to the identification of gene sets or pathways that are behaving differently with respect to the general trend of expression changes across the whole genome (Maciejewski, 2014). GSEA’s approach has led to uncovering interesting biology in the past, as demonstrated by how frequently it has been used and cited.

In the context of RNA-Seq differential analysis, most datasets will be restricted to this option, and thus experiments will be forced to sacrifice knowledge about the absolute copy numbers for an interpretation of the data anchored to whatever the global change is. This should alarm researchers conducting these experiments to recognize the limitations of the current methodologies. This should also push the community to call for technical innovation and standardization that will make more widespread both the use of spike-ins for normalization, and

the validation of reference genes specific to the experiment at hand. Furthermore, this issue regarding the compositional nature of the data is not limited to RNA-Seq, but to many if not all high-throughput ("omics") techniques (see **Appendix 2.4**).

5.2.4 Concerns about the utility of spike-ins

The authors of RUVg (Risso et al., 2014) made two observations that raised concerns about the utility of spike-ins. However, when interpreting the spike-in abundances through the lens of compositional data analysis, the observed behavior of spike-ins is precisely what would be expected (See **Appendix 2.3**). In particular, systematic variation between conditions in the spike-in abundances is expected if there is a global change (**Figure 3.3**; **Figure S2.5**; **Table A2.4**). This was observed in the yeast dataset, with the validated reference gene have a greatly increased abundance in the nitrogen starved cells versus control cells.

However, their other observation raises valid concerns about the current protocol for using spike-ins. They observed a global discrepancy between spike-ins and the rest of the genes when comparing two control libraries to each other (see Figure 4d of (Risso et al., 2014)). This could be partially explained by dropout effects, but it is most likely due to differences in non-poly-adenylated RNA expression (especially rRNA) between the samples. The way that the spike-ins were added in their experiment (adding an equal amount to approximately equal aliquots of the total RNA) causes the spike-ins to also be subject to compositional changes (**Table A2.5**). For bulk RNA-Seq experiments, the standard protocol adds spike-ins to equal amounts of RNA after isolation and selection (poly-A selection or rRNA-depletion), but if there are changes in the excluded RNAs, this protocol impedes the ability of the spike-ins to accurately capture the true fold changes of the RNAs under consideration. In contrast, the approach advocated by Lovén et al (Lovén et al., 2012) was to add spike-ins before RNA isolation, in

proportion to the number of cells. With this approach, the spike-ins can, in principle, accurately capture the behavior of the genes, even when there are non-poly-adenylated RNA changes (**Table A2.6**). There are challenges with using spike-ins in complex tissues (Evans et al., 2018), and there may be technical biases that affect spike-ins differently from endogenous RNAs, but further work must clarify this. What is certain is that future work with spike-ins absolutely must keep in mind the compositional nature of the data being generated, and protocols for bulk RNA-Seq may need to be revised to improve the chance of spike-ins accurately anchoring the data to copy numbers.

5.2.5 Conclusions

In summary, simulating RNA-Seq data using a compositional approach more closely aligns with the kind of data being generated in RNA-Seq. Compositional normalization using negative control features yields a significant improvement over previous methods, in that it performs best in experimental contexts where the composition changes substantially. Importantly, this method can still be safely used in contexts where the compositional changes are unknown. There is much potential to extend the principles of compositional data analysis to other "omics" approaches, since they all generate compositional data; one intriguing possibility is a normalization free method that examines "differential proportionality" (Erb et al., 2018). However, our results from simulation and from real datasets demonstrate that, without access to spike-ins or to validated reference features, a researcher is limited in what conclusions can be drawn from *Seq data because of the compositional nature of the data. This work also makes a strong case for there to be more effort to improve and standardize the use of spike-in controls and validated reference features in all "omics" experiments.

5.3 Compositional Normalization and qPCR: what's next?

It has been known for a quite a while that qPCR requires the use of a validated reference gene to draw valid conclusions. The work presented here is only preliminary and would need additional replication along with some validation with additional techniques (e.g. the NanoString instrument used in (Marguerat et al., 2012)). The key idea is that one or more external RNA is added to the sample in proportion to a unit of biological relevance (most often with in vitro cell line work, it would be number of cells isolated for RNA extraction), and then add at least one more external RNA just before the reverse transcription step to control for reverse transcription and/or qPCR inhibition. This would provide a relatively inexpensive and easy approach to validating any reference gene in any biological context.

5.4 FUS and mitochondrial function: what's next?

Assuming all of the negative data collected in our HEK FUS KO cells is reproducible, the most parsimonious explanation is that FUS serves as an inhibitor of mitochondrial function and biogenesis. Indeed, in our model system, FUS KO cells have more mitochondrial mass and purified mitochondria have *higher* mitochondrial membrane potential, indicating that they are more active than their wild-type counterparts. Work from our lab that was published very recently provides further evidence of this inhibitory role for FUS; in an overexpression model, FUS interacts with Complex V component ATP5B and reduces ATP synthesis. It would be intriguing to know if endogenous FUS can also inhibit ATP synthesis.

Further, all of the mitochondrial-associated genes tested so far are up-regulated after FUS KO (relative to GAPDH), suggesting that FUS inhibits either their transcription, or some downstream step during their processing. If the observation that GAPDH is up-regulated in HEK

cells after FUS KO is reproducible, then it's likely that the genes that were not validated may actually also be increasing as well.

Since all of the work in this thesis was performed in a non-neuronal model, it is unclear what relevance any of this has to neurons. One piece of evidence connecting this work to brain is the observation that FUS KO also reduces mitochondrial size in the brains of mice, a phenocopy of what we observe in the HEK cells. Past work has demonstrated that there is heterogeneity of mitochondrial populations within neurons (Brown et al., 2006; Davey and Clark, 1996; Stauch et al., 2014), and that these have important implications for local mitochondrial function. It will be an intriguing line of future work to explore if FUS has a role to play in regulating this heterogeneity.

5.4 Summary

This work has provided evidence that TDP-43 and FUS both regulate a network of pathways via miRNA-mRNA interactions, and that these networks have biological relevance in both cancer and neurodegeneration. Additionally, this work has created a new normalization approach that will at minimum help researchers more correctly interpret their RNA-Sequencing data, with implications for qPCR and other genome-wide techniques. Finally, this work has provided some additional evidence for the role of FUS in mitochondrial function. All of these represent intriguing and relevant future directions to get one step closer to understanding enough of the pathogenesis of FTD and ALS to develop new diagnostic tools and new therapies.

REFERENCES

Aitchison, J. (2003). The Statistical Analysis of Compositional Data (Blackburn Press).

Aitchison, J. (2008). The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. Proceedings of CoDAWork'08, the 3rd Compositional Data Analysis Workshop.

Al-Chalabi, A., and Hardiman, O. (2013). The epidemiology of ALS: a conspiracy of genes, environment and time. Nature Reviews Neurology *9*, 617–628.

Al-Chalabi, A., Jones, A., Troakes, C., King, A., Al-Sarraj, S., and van den Berg, L.H. (2012). The genetics and neuropathology of amyotrophic lateral sclerosis. Acta Neuropathologica *124*, 339–352.

Al-Chalabi, A., van den Berg, L.H., and Veldink, J. (2017). Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. Nature Reviews Neurology *13*, 96–104.

Al-Shahrour, F., Minguez, P., Tarraga, J., Medina, I., Alloza, E., Montaner, D., and Dopazo, J. (2007a). FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Research *35*, W91–W96.

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Mínguez, P., Montaner, D., and Dopazo, J. (2007b). From genes to functional classes in the study of biological systems. BMC Bioinformatics *8*, 114.

Alami, N.H., Smith, R.B., Carrasco, M.A., Williams, L.A., Winborn, C.S., Han, S.S.W., Kiskinis, E., Winborn, B., Freibaum, B.D., Kanagaraj, A., et al. (2014). Axonal Transport of TDP-43 mRNA Granules Is Impaired by ALS-Causing Mutations. Neuron *81*, 536–543.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600–1607.

Altanbyek, V., Cha, S.-J., Kang, G.-U., Im, D.S., Lee, S., Kim, H.-J., and Kim, K. (2016). Imbalance of mitochondrial dynamics in Drosophila models of amyotrophic lateral sclerosis. Biochemical and Biophysical Research Communications.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology *11*, R106.

Arai, T., Hasegawa, M., Akiyama, H., Ikeda, K., Nonaka, T., Mori, H., Mann, D., Tsuchiya, K., Yoshida, M., Hashizume, Y., et al. (2006). TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Biochemical and Biophysical Research Communications *351*, 602–611.

Archbold, H.C., Jackson, K.L., Arora, A., Weskamp, K., Tank, E.M.H., Li, X., Miguez, R., Dayton, R.D., Tamir, S., Klein, R.L., et al. (2018). TDP43 nuclear export and neurodegeneration in models of amyotrophic lateral sclerosis and frontotemporal dementia. Sci. Rep. *8*, 920.

Ayala, Y.M., De Conti, L., Avendaño-Vázquez, S.E., Dhir, A., Romano, M., D'Ambrogio, A., Tollervey, J., Ule, J., Baralle, M., Buratti, E., et al. (2011). TDP-43 regulates its mRNA levels through a negative feedback loop. Embo J. *30*, 277–288.

Baker, R.G., Hsu, C.J., Lee, D., Jordan, M.S., Maltzman, J.S., Hammer, D.A., Baumgart, T., and Koretzky, G.A. (2009). The Adapter Protein SLP-76 Mediates "Outside-In" Integrin Signaling and Function in T Cells. Molecular and Cellular Biology *29*, 5578–5589.

Baloh, R.H. (2012). How do the RNA-binding proteins TDP-43 and FUS relate to amyotrophic lateral sclerosis and frontotemporal degeneration, and to each other? Current Opinion in Neurology *25*, 701–707.

Bannwarth, S., Ait-El-Mkadem, S., Chaussenot, A., Genin, E.C., Lacas-Gervais, S., Fragaki, K., Berg-Alonso, L., Kageyama, Y., Serre, V., Moore, D.G., et al. (2014). A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. Brain *137*, 2329–2345.

Baran-Gale, J., Kurtz, C.L., Erdos, M.R., Sison, C., Young, A., Fannin, E.E., Chines, P.S., and Sethupathy, P. (2015). Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. Frontiers in Genetics *6*, 352.

Barber, R.D. (2005). GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiological Genomics *21*, 389–395.

Barmada, S.J., Skibinski, G., Korb, E., Rao, E.J., Wu, J.Y., and Finkbeiner, S. (2010). Cytoplasmic mislocalization of TDP-43 is toxic to neurons and enhanced by a mutation associated with familial amyotrophic lateral sclerosis. Journal of Neuroscience *30*, 639–649.

Bartel, D.P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. Cell *136*, 215–233.

Bartel, D.P., and Chen, C.-Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nature Reviews Genetics *5*, 396–400.

Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., and Grant, G.R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature Methods *14*, 135–139.

Bäumer, D., Talbot, K., and Turner, M.R. (2014). Advances in motor neurone disease. J R Soc Med *107*, 14–21.

Beghi, E., Logroscino, G., Chiò, A., Hardiman, O., Mitchell, D., Swingler, R., Traynor, B.J., EURALS Consortium (2006). The epidemiology of ALS and the role of population-based registries. Biochimica Et Biophysica Acta *1762*, 1150–1157.

Ben Langmead, and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359.

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biology *11*, R90.

Bin Zhuo, Emerson, S., Chang, J.H., and Di, Y. (2016). Identifying stably expressed genes from multiple RNA-Seq data sets. PeerJ *4*, e2791.

Blechingberg, J., Luo, Y., Bolund, L., Damgaard, C.K., and Nielsen, A.L. (2012). Gene Expression Responses to FUS, EWS, and TAF15 Reduction and Stress Granule Sequestration Analyses Identifies FET-Protein Non-Redundant Functions. PLoS ONE *7*, e46251.

Bohovych, I., and Khalimonchuk, O. (2016). Sending Out an SOS: Mitochondria as a Signaling Hub. Front. Cell Dev. Biol. *4*, 524.

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

Bottomly, D., Walter, N.A.R., Hunter, J.E., Darakjian, P., Kawane, S., Buck, K.J., Searles, R.P., Mooney, M., McWeeney, S.K., and Hitzemann, R. (2011). Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. PLoS ONE *6*, e17820.

Braak, H., Ludolph, A.C., Neumann, M., Ravits, J., and Del Tredici, K. (2017). Pathological TDP-43 changes in Betz cells differ from those in bulbar and spinal α-motoneurons in sporadic amyotrophic lateral sclerosis. Acta Neuropathologica *133*, 79–90.

Brand, M.D., and Nicholls, D.G. (2011). Assessing mitochondrial dysfunction in cells. Biochem. J. *435*, 297–312.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology *34*, 525–527.

Brockington, A., Ning, K., Heath, P.R., Wood, E., Kirby, J., Fusi, N., Lawrence, N., Wharton, S.B., Ince, P.G., and Shaw, P.J. (2012). Unravelling the enigma of selective vulnerability in neurodegeneration: motor neurons resistant to degeneration in ALS show distinct gene expression characteristics and decreased susceptibility to excitotoxicity. Acta Neuropathologica *125*, 95–109.

Brown, M.R., Sullivan, P.G., and Geddes, J.W. (2006). Synaptic mitochondria are more susceptible to Ca2+overload than nonsynaptic mitochondria. J. Biol. Chem. *281*, 11658–11668.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Buratti, E., and Baralle, F.E. (2001). Characterization and Functional Implications of the RNA Binding Properties of Nuclear Factor TDP-43, a Novel Splicing Regulator of CFTR Exon 9. Journal of Biological Chemistry *276*, 36337–36343.

Buratti, E. (2015). Functional Significance of TDP-43 Mutations in Disease. Adv. Genet. *91*, 1–53.

Buratti, E., De Conti, L., Stuani, C., Romano, M., Baralle, M., and Baralle, F. (2010). Nuclear factor TDP-43 can affect selected microRNA levels. FEBS Journal *277*, 2268–2281.

Buratti, E., Romano, M., and Baralle, F.E. (2013). TDP-43 high throughput screening analyses in neurodegeneration: Advantages and pitfalls. Mol. Cell. Neurosci. *56*, 465–474.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009). The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry *55*, 611–622.

Camats, M., Guil, S., Kokolo, M., and Bach-Elias, M. (2008). P68 RNA helicase (DDX5) alters activity of cis-and trans-acting factors of the alternative splicing of H-Ras. PLoS ONE *3*, e2926.

Campos-Melo, D., Droppelmann, C.A., He, Z., Volkening, K., and Strong, M.J. (2013). Altered microRNA expression profile in amyotrophic lateral sclerosis: a role in the regulation of NFL mRNA levels. Mol Brain *6*, 26.

Campos-Melo, D., Droppelmann, C.A., Volkening, K., and Strong, M.J. RNA-binding proteins as molecular links between cancer and neurodegeneration. Biogerontology *15*, 587–610.

Carrí, M.T., Valle, C., Bozzo, F., and Cozzolino, M. (2015). Oxidative stress and mitochondrial damage: importance in non-SOD1 ALS. Front. Cell. Neurosci. *9*, 41.

Casafont, I., Bengoechea, R., Tapia, O., Berciano, M.T., and Lafarga, M. (2009). TDP-43 localizes in mRNA transcription and processing sites in mammalian neurons. Journal of Structural Biology *167*, 235–241.

Chandel, N.S. (2014). Mitochondria as signaling organelles. BMC Biology *12*, 34.

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., and Tyler, J.K. (2015). The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. Molecular and Cellular Biology *36*, 662–667.

Chen, X., Fan, Z., McGee, W., Chen, M., Kong, R., Wen, P., Xiao, T., Chen, X., Liu, J., Zhu, L., et al. (2018). TDP-43 regulates cancer-associated microRNAs. Protein & Cell *9*, 848–866.

Chen, Y., Deng, J., Wang, P., Yang, M., Chen, X., Zhu, L., Liu, J., Lu, B., Shen, Y., Fushimi, K., et al. (2016). PINK1 and Parkin are genetic modifiers for FUS-induced neurodegeneration. Human Molecular Genetics *25*, 5059–5068.

Chen, Y., Yang, M., Deng, J., Chen, X., Ye, Y., Zhu, L., Liu, J., Ye, H., Shen, Y., Li, Y., et al. (2011). Expression of human FUS protein in Drosophila leads to progressive neurodegeneration. Protein & Cell *2*, 477–486.

Chiang, P.-M., Ling, J., Jeong, Y.H., Price, D.L., Aja, S.M., and Wong, P.C. (2010). Deletion of TDP-43 down-regulates Tbc1d1, a gene linked to obesity, and alters body fat metabolism. Proceedings of the National Academy of Sciences *107*, 16320–16324.

Chiò, A., Traynor, B.J., Lombardo, F., Fimognari, M., Calvo, A., Ghiglione, P., Mutani, R., and Restagno, G. (2008). Prevalence of SOD1 mutations in the Italian ALS population. Neurology *70*, 533–537.

Collins, F.S., and Barker, A.D. (2007). Mapping the cancer genome. Sci. Am. *296*, 50–57.

Colombrita, C., Onesto, E., Buratti, E., La Grange, De, P., Gumina, V., Baralle, F.E., Silani, V., and Ratti, A. (2015). From transcriptomic to protein level changes in TDP-43 and FUS loss-of-function cell models. Biochimica Et Biophysica Acta *1849*, 1398-1410.

Colombrita, C., Onesto, E., Megiorni, F., Pizzuti, A., Baralle, F.E., Buratti, E., Silani, V., and Ratti, A. (2012). TDP-43 and FUS RNA-binding Proteins Bind Distinct Sets of Cytoplasmic Messenger RNAs and Differently Regulate Their Post-transcriptional Fate in Motoneuron-like Cells. Journal of Biological Chemistry *287*, 15635–15647.

Cooper-Knock, J., Robins, H., Niedermoser, I., Wyles, M., Heath, P.R., Higginbottom, A., Walsh, T., Kazoka, M., Consortium, P.M.A.S., Ince, P.G., et al. (2017). Targeted Genetic Screen in Amyotrophic Lateral Sclerosis Reveals Novel Genetic Variants with Synergistic Effect on Clinical Phenotype. Front. Mol. Neurosci. *10*, 146.

Cottet Rousselle, C., Ronot, X., Leverve, X., and Mayol, J.F. (2011). Cytometric assessment of mitochondria using fluorescent probes. Cytometry Part A *79A*, 405–425.

Cozzolino, M., and Carrí, M.T. (2012). Mitochondrial dysfunction in ALS. Progress in Neurobiology *97*, 54–66.

Crozat, A., Åman, P., Mandahl, N., and Ron, D. (1993). Fusion of CHOP to a novel RNA-binding protein in human myxoid liposarcoma. Nature *363*, 640–644.

Czech, B., and Hannon, G.J. (2011). Small RNA sorting: matchmaking for Argonautes. Nature Reviews Genetics *12*, 19–31.

Davey, G.P., and Clark, J.B. (1996). Threshold Effects and Control of Oxidative Phosphorylation in Nonsynaptic Rat Brain Mitochondria. Journal of Neurochemistry *66*, 1617–1624.

Davey, G.P., Peuchen, S., and Clark, J.B. (1998). Energy Thresholds in Brain Mitochondria. J. Biol. Chem. *273*, 12753–12757.

Davis, S.A., Itaman, S., Khalid-Janney, C.M., Sherard, J.A., Dowell, J.A., Cairns, N.J., and Gitcho, M.A. (2018). TDP-43 interacts with mitochondrial proteins critical for mitophagy and mitochondrial dynamics. Neuroscience Letters *678*, 8–15.

De Giorgio, F., Maduro, C., Fisher, E.M.C., and Acevedo Arozena, A. (2019). Transgenic and physiological mouse models give insights into different aspects of amyotrophic lateral sclerosis. Disease Models & Mechanisms *12*, dmm037424.

Deng, J., Wang, P., Chen, X., Cheng, H., Liu, J., Fushimi, K., Zhu, L., and Wu, J.Y. (2018). FUS interacts with ATP synthase beta subunit and induces mitochondrial unfolded protein response in cellular and animal models. Proceedings of the National Academy of Sciences *115*, E9678–E9686.

Deng, J., Yang, M., Chen, Y., Chen, X., Liu, J., Sun, S., Cheng, H., Li, Y., Bigio, E., Mesulam, M., et al. (2015). FUS Interacts with HSP60 to Promote Mitochondrial Damage. PLoS Genetics *11*, e1005357.

Di Carlo, V., Grossi, E., Laneve, P., Morlando, M., Dini Modigliani, S., Ballarino, M., Bozzoni, I., and Caffarelli, E. (2013). TDP-43 Regulates the Microprocessor Complex Activity During In Vitro Neuronal Differentiation. *48*, 952–963.

Dini Modigliani, S., Morlando, M., Errichelli, L., Sabatelli, M., and Bozzoni, I. (2014). An ALS-associated mutation in the FUS 3′-UTR disrupts a microRNA–FUS regulatory circuitry. Nature Communications *5*.

Dormann, D., Rodde, R., Edbauer, D., Bentmann, E., Fischer, I., Hruscha, A., Than, M.E., Mackenzie, I.R.A., Capell, A., Schmid, B., et al. (2010). ALS-associated fused in sarcoma (FUS) mutations disrupt Transportin-mediated nuclear import. Embo J. *29*, 2841–2857.

D'Alton, S., Altshuler, M., Cannon, A., Dickson, D.W., Petrucelli, L., and Lewis, J. (2014). Divergent Phenotypes in Mutant TDP-43 Transgenic Mice Highlight Potential Confounds in TDP-43 Transgenic Modeling. PLoS ONE *9*, e86513.

Ederle, H., Funk, C., Abou-Ajram, C., Hutten, S., Funk, E.B.E., Kehlenbach, R.H., Bailer, S.M., and Dormann, D. (2018). Nuclear egress of TDP-43 and FUS occurs independently of Exportin-1/CRM1. Sci. Rep. *8*, 7084.

Efimova, A.D., Ovchinnikov, R.K., Roman, A.Y., Maltsev, A.V., Grigoriev, V.V., Kovrazhkina, E.A., and Skvortsova, V.I. (2017). The FUS protein: Physiological functions and a role in amyotrophic lateral sclerosis. Mol Biol *51*, 341–351.

Eitan, C., and Hornstein, E. (2016). Vulnerability of microRNA biogenesis in FTD–ALS. Brain Research *1647*, 105–111.

Ejigu, B.A., Valkenborg, D., Baggerman, G., Vanaerschot, M., Witters, E., Dujardin, J.-C., Burzykowski, T., and Berg, M. (2013). Evaluation of Normalization Methods to Pave the Way Towards Large-Scale LC-MS-Based Metabolomics Profiling Experiments. OMICS: a Journal of Integrative Biology *17*, 473–485.

Emde, A., Eitan, C., Liou, L.L., Libby, R.T., Rivkin, N., Magen, I., Reichenstein, I., Oppenheim, H., Eilam, R., Silvestroni, A., et al. (2015). Dysregulated miRNA biogenesis downstream of cellular stress and ALS-causing mutations: a new mechanism for ALS. Embo J. *34*, 2633–2651.

Erb, I., Quinn, T., Lovell, D., and Notredame, C. (2018). Differential Proportionality - A Normalization-Free Approach To Differential Gene Expression. bioRxiv, 10.1101/134536.

Erkkinen, M.G., Kim, M.-O., and Geschwind, M.D. (2018). Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases. Cold Spring Harbor Perspectives in Biology *10*, a033118.

Evans, C., Hardin, J., and Stoebel, D.M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics *19*, 776–792.

Fang, H.Y., Chen, S.B., Guo, D.J., Pan, S.Y., and Yu, Z.L. (2011). Proteomic identification of differentially expressed proteins in curcumin-treated MCF-7 cells. Phytomedicine *18*, 697–703.

Fasold, M., and Binder, H. (2014). Variation of RNA Quality and Quantity Are Major Sources of Batch Effects in Microarray Expression Data. Microarrays *3*, 322–339.

Feiguin, F., Godena, V.K., Romano, G., D'Ambrogio, A., Klima, R., and Baralle, F.E. (2009). Depletion of TDP-43 affects Drosophila motoneurons terminal synapsis and locomotive behavior. FEBS Letters *583*, 1586–1592.

Fernandes, A.M.G.C. (2012). MicroRNA regulation by the DNA and RNA binding proteins EWS and FUS. PhD Thesis, University of Lisbon.

Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2014 2:1 *2*, 15.

Ferrari, R., Forabosco, P., Vandrovcova, J., Botía, J.A., Guelfi, S., Warren, J.D., UK Brain Expression Consortium (UKBEC), Momeni, P., Weale, M.E., Ryten, M., et al. (2016). Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. Mol Neurodegeneration *11*, 1615.

Ferreira, T., Wilson, S.R., Choi, Y.G., Risso, D., Dudoit, S., Speed, T.P., and Ngai, J. (2014). Silencing of Odorant Receptor Genes by G Protein βγ Signaling Ensures the Expression of One Odorant Receptor per Olfactory Sensory Neuron. Neuron *81*, 847–859.

Figueroa-Romero, C., Hur, J., Lunn, J.S., Paez-Colasante, X., Bender, D.E., Yung, R., Sakowski, S.A., and Feldman, E.L. (2016). Expression of microRNAs in human post-mortem amyotrophic lateral sclerosis spinal cords provides insight into disease mechanisms. Molecular and Cellular Neuroscience *71*, 34–45.

Finnegan, E.F., and Pasquinelli, A.E. (2013). MicroRNA biogenesis: regulating the regulators. Critical Reviews in Biochemistry and Molecular Biology *48*, 51–68.

Fogarty, M.J. (2018). Driven to decay: Excitability and synaptic abnormalities in amyotrophic lateral sclerosis. Brain Research Bulletin *140*, 318–333.

Fontana, F., Siva, K., and Denti, M.A. (2015). A network of RNA and protein interactions in Fronto Temporal Dementia. Front. Mol. Neurosci. *8*, 9.

Frazee, A.C., Jaffe, A.E., Ben Langmead, and Leek, J.T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics *17*, 2778–2784.

Freischmidt, A., Müller, K., Ludolph, A.C., and Weishaupt, J.H. (2013). Systemic dysregulation of TDP-43 binding microRNAs in amyotrophic lateral sclerosis. Acta Neuropathol Commun *1*, 42–42.

Fujii, R., Okabe, S., Urushido, T., Inoue, K., Yoshimura, A., Tachibana, T., Nishikawa, T., Hicks, G.G., and Takumi, T. (2005). The RNA binding protein TLS is translocated to dendritic spines by mGluR5 activation and regulates spine morphology. Curr. Biol. *15*, 587–593.

Fujioka, Y., Ishigaki, S., Masuda, A., Iguchi, Y., Udagawa, T., Watanabe, H., Katsuno, M., Ohno, K., and Sobue, G. (2013). FUS-regulated region- and cell-type-specific transcriptome is associated with cell selectivity in ALS/FTLD. Sci. Rep. *3*, 1–12.

Furukawa, Y., Suzuki, Y., Fukuoka, M., Nagasawa, K., Nakagome, K., Shimizu, H., Mukaiyama, A., and Akiyama, S. (2016). A molecular mechanism realizing sequence-specific recognition of nucleic acids by TDP-43. Sci. Rep. *6*, 20576.

Fushimi, K., Long, C., Jayaram, N., Chen, X., Li, L., and Wu, J.Y. (2011). Expression of human FUS/TLS in yeast leads to protein aggregation and cytotoxicity, recapitulating key features of FUS proteinopathy. Protein & Cell *2*, 141–149.

Galvin, J.E., Howard, D.H., Denny, S.S., Dickinson, S., and Tatton, N. (2017). The social and economic burden of frontotemporal degeneration. Neurology *89*, 2049–2056.

Gama-Carvalho, M., L Garcia-Vaquero, M., R Pinto, F., Besse, F., Weis, J., Voigt, A., Schulz, J.B., and Las Rivas, De, J. (2017). Linking amyotrophic lateral sclerosis and spinal muscular atrophy through RNA-transcriptome homeostasis: a genomics perspective. Journal of Neurochemistry *141*, 12–30.

Gao, F.-B., Almeida, S., and Lopez Gonzalez, R. (2017). Dysregulated molecular pathways in amyotrophic lateral sclerosis-frontotemporal dementia spectrum disorder. Embo J *36*, 2931–2950.

Gascon, E., and Gao, F.-B. (2012). Cause or Effect: Misregulation of microRNA Pathways in Neurodegeneration. Frontiers in Genetics *6*, 48.

Gascon, E., and Gao, F.-B. (2014). The Emerging Roles of MicroRNAs in the Pathogenesis of Frontotemporal Dementia–Amyotrophic Lateral Sclerosis (FTD-ALS) Spectrum Disorders. J Neurogenet *28*, 30–40.

Geser, F., Lee, V.M.Y., and Trojanowski, J.Q. (2010). Amyotrophic lateral sclerosis and frontotemporal lobar degeneration: A spectrum of TDP-43 proteinopathies. Neuropathology *30*, 103–112.

Geser, F., Martinez-Lage, M., Robinson, J., Uryu, K., Neumann, M., Brandmeir, N.J., Xie, S.X., Kwong, L.K., Elman, L., McCluskey, L., et al. (2009). Clinical and pathological continuum of multisystem TDP-43 proteinopathies. Arch. Neurol. *66*, 180–189.

Gibson, S.B., Abbott, D., Farnham, J.M., Thai, K.K., McLean, H., Figueroa, K.P., Bromberg, M.B., Pulst, S.M., and Cannon-Albright, L. (2016). Population-based risks for cancer in patients with ALS. Neurology *87*, 289–294.

Gitcho, M.A., Bigio, E.H., Mishra, M., Johnson, N., Weintraub, S., Mesulam, M., Rademakers, R., Chakraverty, S., Cruchaga, C., Morris, J.C., et al. (2009). TARDBP 3′-UTR variant in autopsy-confirmed frontotemporal lobar degeneration with TDP-43 proteinopathy. Acta Neuropathol *118*, 633.

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. *8*, 57.

Gregory, R.I., Yan, K.-P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235–240.

Guo, L., and Shorter, J. (2017). Biology and Pathobiology of TDP-43 and Emergent Therapeutic Strategies. Cold Spring Harbor Perspectives in Medicine *7*, a024554.

Guo, W., Chen, Y., Zhou, X., Kar, A., Ray, P., Chen, X., Rao, E.J., Yang, M., Ye, H., Zhu, L., et al. (2011). An ALS-associated mutation affecting TDP-43 enhances protein aggregation, fibril formation and neurotoxicity. Nature Structural & Molecular Biology *18*, 822–830.

Guo, W., Fumagalli, L., Prior, R., and Van Den Bosch, L. (2017). Current Advances and Limitations in Modeling ALS/FTD in a Dish Using Induced Pluripotent Stem Cells. Front. Neurosci. *11*, 1282.

Haramati, S., Chapnik, E., Sztainberg, Y., Eilam, R., Zwang, R., Gershoni, N., McGlinn, E., Heiser, P.W., Wills, A.-M., Wirguin, I., et al. (2010). miRNA malfunction causes spinal motor neuron disease. Proceedings of the National Academy of Sciences *107*, 13111–13116.

Hardiman, O., Al-Chalabi, A., Chiò, A., Corr, E.M., Logroscino, G., Robberecht, W., Shaw, P.J., Simmons, Z., and van den Berg, L.H. (2017). Amyotrophic lateral sclerosis. Nature Reviews Disease Primers 2017 3 *3*, 17071.

Hardwick, S.A., Deveson, I.W., and Mercer, T.R. (2017). Reference standards for next-generation sequencing. Nature Reviews Genetics *18*, 473–484.

Hart, S.N., Therneau, T.M., Zhang, Y., Poland, G.A., and Kocher, J.-P. (2013). Calculating Sample Size Estimates for RNA Sequencing Data. Journal of Computational Biology *20*, 970–978.

Hawkins, R.D., Hon, G.C., and Ren, B. (2010). Next-generation genomics: an integrative approach. Nature Reviews Genetics *11*, 476-486.

Hicks, G.G., Singh, N., Nashabi, A., Mai, S., Bozek, G., Klewes, L., Arapovic, D., White, E.K., Koury, M.J., Oltz, E.M., et al. (2000). Fus deficiency in mice results in defective B-lymphocyte development and activation, high levels of chromosomal instability and perinatal death. Nature Genetics *24*, 175–179.

Highley, J.R., Kirby, J., Jansweijer, J.A., Webb, P.S., Hewamadduma, C.A., Heath, P.R., Higginbottom, A., Raman, R., Ferraiuolo, L., Cooper-Knock, J., et al. (2014). Loss of nuclear TDP-43 in amyotrophic lateral sclerosis (ALS) causes altered expression of splicing machinery and widespread dysregulation of RNA splicing in motor neurones. Neuropathology and Applied Neurobiology *40*, 670–685.

Hill, S.J., Mordes, D.A., Cameron, L.A., Neuberg, D.S., Landini, S., Eggan, K., and Livingston, D.M. (2016). Two familial ALS proteins function in prevention/repair of transcription-associated DNA damage. Proceedings of the National Academy of Sciences *113*, E7701–E7709.

Hoell, J.I., Larsson, E., Runge, S., Nusbaum, J.D., Duggimpudi, S., Farazi, T.A., Hafner, M., Borkhardt, A., Sander, C., and Tuschl, T. (2011). RNA targets of wild-type and mutant FET family proteins. Nat Struct Mol Biol *18*, 1428–1431.

Honda, D., Ishigaki, S., Iguchi, Y., Fujioka, Y., Udagawa, T., Masuda, A., Ohno, K., Katsuno, M., and Sobue, G. (2014). The ALS/FTLD-related RNA-binding proteins TDP-43 and FUS have common downstream RNA targets in cortical neurons. FEBS Open Bio *4*, 1–10.

Huang, C., Zhou, H., Tong, J., Chen, H., Liu, Y.-J., Wang, D., Wei, X., and Xia, X.-G. (2011). FUS Transgenic Rats Develop the Phenotypes of Amyotrophic Lateral Sclerosis and Frontotemporal Lobar Degeneration. PLoS Genetics *7*, e1002011.

Huang, E.J., Zhang, J., Geser, F., Trojanowski, J.Q., Strober, J.B., Dickson, D.W., Brown, R.H., Jr, Shapiro, B.E., and Lomen-Hoerth, C. (2010). Extensive FUS-Immunoreactive Pathology in

Juvenile Amyotrophic Lateral Sclerosis with Basophilic Inclusions. Brain Pathology *20*, 1069–1076.

Huang, H.-C., Niu, Y., and Qin, L.-X. (2015). Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. Cancer Informatics *14*, 57–67.

Huang, P., Yu, T., and Yoon, Y. (2007). Mitochondrial clustering induced by overexpression of the mitochondrial fusion protein Mfn2 causes mitochondrial dysfunction and cell death. Eur. J. Cell Biol. *86*, 289–302.

Hunt, R.J., and Bateman, J.M. (2018). Mitochondrial retrograde signaling in the nervous system. FEBS Letters *592*, 663–678.

Ishigaki, S., and Sobue, G. (2018). Importance of Functional Loss of FUS in FTLD/ALS. Front. Mol. Biosci. *5*, 275.

Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., Urano, F., Sobue, G., and Ohno, K. (2012). Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. Sci. Rep. *2*, 529.

Izumikawa, K., Nobe, Y., Yoshikawa, H., Ishikawa, H., Miura, Y., Nakayama, H., Nonaka, T., Hasegawa, M., Egawa, N., Inoue, H., et al. (2017). TDP-43 stabilises the processing intermediates of mitochondrial transcripts. Sci. Rep. *7*, 7709.

Jawaid, A., Khan, R., Polymenidou, M., and Schulz, P.E. (2018). Disease-modifying effects of metabolic perturbations in ALS/FTLD. Mol Neurodegeneration *13*, 63.

Jia, C., Hu, Y., Kelly, D., Kim, J., Li, M., and Zhang, N.R. (2017). Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Research *45*, 10978–10988.

Jiang, L.-L., Xue, W., Hong, J.-Y., Zhang, J.-T., Li, M.-J., Yu, S.-N., He, J.-H., and Hu, H.-Y. (2017). The N-terminal dimerization is required for TDP-43 splicing activity. Sci. Rep. *7*, 6196.

Jiang, Z., Wang, W., Perry, G., Zhu, X., and Wang, X. (2015). Mitochondrial dynamic abnormalities in amyotrophic lateral sclerosis. Translational Neurodegeneration *4*, 295.

Josephs, K.A., Whitwell, J.L., Parisi, J.E., Petersen, R.C., Boeve, B.F., Jack, C.R., Jr, and Dickson, D.W. (2010). Caudate atrophy on MRI is a characteristic feature of FTLD-FUS. Eur J Neurol *17*, 969–975.

Kaivorinne, A.-L. (2012). Frontotemporal Lobar Degeneration in Finland: Molecular genetics and clinical aspects. PhD Thesis, University of Oulu.

Kansal, K., Mareddy, M., Sloane, K.L., Minc, A.A., Rabins, P.V., McGready, J.B., and Onyike, C.U. (2016). Survival in Frontotemporal Dementia Phenotypes: A Meta-Analysis. Dementia and Geriatric Cognitive Disorders *41*, 109–122.

Kapeli, K., Pratt, G.A., Vu, A.Q., Hutt, K.R., Martinez, F.J., Sundararaman, B., Batra, R., Freese, P., Lambert, N.J., Huelga, S.C., et al. (2016). Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. Nature Communications *7*, 12143.

Karres, J.S., Hilgers, V., Carrera, I., Treisman, J., and Cohen, S.M. (2007). The Conserved microRNA MiR-8 Tunes Atrophin Levels to Prevent Neurodegeneration in Drosophila. Cell *131*, 136–145.

Katisko, K., Haapasalo, A., Koivisto, A., Krüger, J., Hartikainen, P., Korhonen, V., Helisalmi, S., Herukka, S.-K., Remes, A.M., and Solje, E. (2018). Low Prevalence of Cancer in Patients with Frontotemporal Lobar Degeneration. Journal of Alzheimer's Disease *62*, 789–794.

Kawahara, Y., and Mieda-Sato, A. (2012). TDP-43 promotes microRNA biogenesis as a component of the Drosha and Dicer complexes. Proceedings of the National Academy of Sciences *109*, 3347–3352.

Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nature Methods *11*, 740–742.

Kim, K.Y., Lee, H.-W., Shim, Y.-M., Mook-Jung, I., Jeon, G.S., and Sung, J.-J. (2015). A phosphomimetic mutant TDP-43 (S409/410E) induces Drosha instability and cytotoxicity in Neuro 2A cells. Biochemical and Biophysical Research Communications *464*, 236–243.

King, I.N., Yartseva, V., Salas, D., Kumar, A., Heidersbach, A., Ando, D.M., Stallings, N.R., Elliott, J.L., Srivastava, D., and Ivey, K.N. (2014). The RNA-binding Protein TDP-43 Selectively Disrupts MicroRNA-1/206 Incorporation into the RNA-induced Silencing Complex. Journal of Biological Chemistry *289*, 14272–14272.

Kino, Y., Washizu, C., Aquilanti, E., Okuno, M., Kurosawa, M., Yamada, M., Doi, H., and Nukina, N. (2011). Intracellular localization and splicing regulation of FUS/TLS are variably affected by amyotrophic lateral sclerosis-linked mutations. Nucleic Acids Research *39*, 2781–2798.

Kino, Y., Washizu, C., Kurosawa, M., Yamada, M., Miyazaki, H., Akagi, T., Hashikawa, T., Doi, H., Takumi, T., Hicks, G.G., et al. (2015). FUS/TLS deficiency causes behavioral and pathological abnormalities distinct from amyotrophic lateral sclerosis. Acta Neuropathol Commun *3*, 24.

Knopman, D.S., and Roberts, R.O. (2011). Estimating the Number of Persons with Frontotemporal Lobar Degeneration in the US Population. J Mol Neurosci *45*, 330–335.

Kocerha, J., Kouri, N., Baker, M., Finch, N., DeJesus-Hernandez, M., Gonzalez, J., Chidamparam, K., Josephs, K.A., Boeve, B.F., Graff-Radford, N.R., et al. (2011). Altered microRNA expression in frontotemporal lobar degeneration with TDP-43 pathology caused by progranulin mutations. BMC Genomics *12*, 527.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research *42*, D68–D73.

Kraemer, B.C., Schuck, T., Wheeler, J.M., Robinson, L.C., Trojanowski, J.Q., Lee, V.M.Y., and Schellenberg, G.D. (2010). Loss of murine TDP-43 disrupts motor function and plays an essential role in embryogenesis. Acta Neuropathologica *119*, 409–419.

Krug, L., Chatterjee, N., Borges-Monroy, R., Hearn, S., Liao, W.-W., 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. PLoS Genetics *13*, e1006635.

Krzywinski, M., Birol, I., Jones, S.J.M., and Marra, M.A. (2012). Hive plots-rational approach to visualizing networks. Briefings in Bioinformatics *13*, 627–644.

Kulkarni, M.M. (2001). Digital Multiplexed Gene Expression Analysis Using the NanoString nCounter System. Current Protocols in Molecular Biology *326*, 25B.10.1-25B.10.17.

Kuo, P.-H., Chiang, C.-H., Wang, Y.-T., Doudeva, L.G., and Yuan, H.S. (2014). The crystal structure of TDP-43 RRM1-DNA complex reveals the specific recognition for UG- and TG-rich nucleic acids. Nucleic Acids Research *42*, 4712–4722.

Kuroda, M., Sok, J., Webb, L., Baechtold, H., Urano, F., Yin, Y., Chung, P., de Rooij, D.G., Akhmedov, A., Ashley, T., et al. (2000). Male sterility and enhanced radiation sensitivity in TLS−/− mice. Embo J *19*, 453–462.

Kwiatkowski, T.J., Bosco, D.A., LeClerc, A.L., Tamrazian, E., Vanderburg, C.R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E.J., Munsat, T., et al. (2009). Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. Science *323*, 1205–1208.

Lagier-Tourenne, C., Polymenidou, M., and Cleveland, D.W. (2010). TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. Human Molecular Genetics *19*, R46–R64.

Lagier-Tourenne, C., Polymenidou, M., Hutt, K.R., Vu, A.Q., Baughn, M., Huelga, S.C., Clutario, K.M., Ling, S.-C., Liang, T.Y., Mazur, C., et al. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. Nature Neuroscience *15*, 1488–1497.

Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., et al. (2014). IVT-seq reveals extreme bias in RNA sequencing. Genome Biology *15*, R86.

Larkindale, J., Yang, W., Hogan, P.F., Simon, C.J., Zhang, Y., Jain, A., Habeeb-Louks, E.M., Kennedy, A., and Cwik, V.A. (2014). Cost of illness for neuromuscular diseases in the United States. Muscle Nerve *49*, 431–438.

Lau, D.H.W., Hartopp, N., Welsh, N.J., Mueller, S., Glennon, E.B., Mórotz, G.M., Annibali, A., Gomez Suaga, P., Stoica, R., Paillusson, S., et al. (2018). Disruption of ER−mitochondria signalling in fronto-temporal dementia and related amyotrophic lateral sclerosis. Cell Death & Disease *9*, 327.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology *15*, R29.

Le, T.D., Liu, L., Zhang, J., Liu, B., and Li, J. (2015). From miRNA regulation to miRNA-TF co-regulation: computational approaches and challenges. Briefings in Bioinformatics *16*, 475-496.

Leshkowitz, D., Horn-Saban, S., Parmet, Y., and Feldmesser, E. (2013). Differences in microRNA detection levels are technology and sequence dependent. RNA *19*, 527–538.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 1.

Li, W., Jin, Y., Prazak, L., Hammell, M., and Dubnau, J. (2012). Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. PLoS ONE *7*, e44099.

Li, Y., Ray, P., Rao, E.J., Shi, C., Guo, W., Chen, X., Woodruff, E.A., III, Fushimi, K., and Wu, J.Y. (2010). A Drosophila model for TDP-43 proteinopathy. Proceedings of the National Academy of Sciences *107*, 3169–3174.

Li, Y., Liang, C., Wong, K.-C., Jin, K., and Zhang, Z. (2014). Inferring probabilistic miRNA–mRNA interaction signatures in cancers: a role-switch approach. Nucleic Acids Research *42*, e76.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional Amplification in Tumor Cells with Elevated c-Myc. Cell *151*, 56–67.

Lindström, M., and Liu, B. (2018). Yeast as a Model to Unravel Mechanisms Behind FUS Toxicity in Amyotrophic Lateral Sclerosis. Front. Mol. Neurosci. *11*, 37.

Ling, J.P., Pletnikova, O., Troncoso, J.C., and Wong, P.C. (2015). TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science *349*, 650–655.

Ling, S.-C. (2018). Synaptic Paths to Neurodegeneration: The Emerging Role of TDP-43 and FUS in Synaptic Functions. Neural Plasticity *2018*, 1–13.

Liu, X., Niu, C., Ren, J., Zhang, J., Xie, X., Zhu, H., Feng, W., and Gong, W. (2013). The RRM domain of human fused in sarcoma protein reveals a non-canonical nucleic acid binding site. Biochimica Et Biophysica Acta *1832*, 375–385.

Lomen-Hoerth, C. (2011). Clinical Phenomenology and Neuroimaging Correlates in ALS-FTD. J Mol Neurosci *45*, 656–662.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 31.

Lovell, D., Müller, W., Taylor, J., Zwart, A., and Helliwell, C. (2011). Proportions, Percentages, PPM: Do the Molecular Biosciences Treat Compositional Data Right? In Compositional Data Analysis, V. Pawlowsky-Glahn, and A. Buccianti, eds. (Chichester, UK: John Wiley & Sons, Ltd), pp. 191–207.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., and Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. PLoS Computational Biology *11*, e1004075.

Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., and Young, R.A. (2012). Revisiting Global Gene Expression Analysis. Cell *151*, 476–482.

Lukavsky, P.J., Daujotyte, D., Tollervey, J.R., Ule, J., Stuani, C., Buratti, E., Baralle, F.E., Damberger, F.F., and Allain, F.H.-T. (2013). Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. Nat Struct Mol Biol *20*, 1443–1449.

Luxenhofer, G., Helmbrecht, M.S., Langhoff, J., Giusti, S.A., Refojo, D., and Huber, A.B. (2014). MicroRNA-9 promotes the switch from early-born to late-born motor neuron populations by regulating Onecut transcription factor expression. Developmental Biology *386*, 358–370.

Maciejewski, H. (2014). Gene set analysis methods: statistical models and methodological differences. Briefings in Bioinformatics *15*, 504–518.

Mackenzie, I.R.A., and Neumann, M. (2016). Molecular neuropathology of frontotemporal dementia: insights into disease mechanisms from postmortem studies. Journal of Neurochemistry *138*, 54–70.

Mackenzie, I.R.A., and Neumann, M. (2017a). Fused in Sarcoma Neuropathology in Neurodegenerative Disease. Cold Spring Harbor Perspectives in Medicine *7*, a024299.

Mackenzie, I.R.A., Bigio, E.H., Ince, P.G., Geser, F., Neumann, M., Cairns, N.J., Kwong, L.K., Forman, M.S., Ravits, J., Stewart, H., et al. (2007). Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. Ann Neurol. *61*, 427–434.

Mackenzie, I.R.A., Neumann, M., Baborie, A., Sampathu, D.M., Plessis, D., Jaros, E., Perry, R.H., Trojanowski, J.Q., Mann, D.M.A., and Lee, V.M.Y. (2011). A harmonized classification system for FTLD-TDP pathology. Acta Neuropathologica *122*, 111–113.

Mackenzie, I.R.A., Neumann, M., Bigio, E.H., Cairns, N.J., Alafuzoff, I., Kril, J., Kovacs, G.G., Ghetti, B., Halliday, G., Holm, I.E., et al. (2008). Nomenclature for neuropathologic subtypes of

frontotemporal lobar degeneration: consensus recommendations. Acta Neuropathologica *117*, 15–18.

Mackenzie, I.R.A., Neumann, M., Bigio, E.H., Cairns, N.J., Alafuzoff, I., Kril, J., Kovacs, G.G., Ghetti, B., Halliday, G., Holm, I.E., et al. (2009). Nomenclature and nosology for neuropathologic subtypes of frontotemporal lobar degeneration: an update. Acta Neuropathologica *119*, 1–4.

Mackenzie, I.R., and Neumann, M. (2017b). Reappraisal of TDP-43 pathology in FTLD-U subtypes. Acta Neuropathologica *134*, 79–96.

Mackenzie, I.R., Rademakers, R., and Neumann, M. (2010). TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. Lancet Neurol *9*, 995–1007.

Mancarci, B.O., Toker, L., Tripathy, S.J., Li, B., Rocco, B., Sibille, E., and Pavlidis, P. (2017). Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. eNeuro *4*, e0212–17.2017.

Mao, Y., Kuo, S.-W., Le Chen, Heckman, C.J., and Jiang, M.C. (2017). The essential and downstream common proteins of amyotrophic lateral sclerosis: A protein-protein interaction network analysis. PLoS ONE *12*, e0172246.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bähler, J. (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. Cell *151*, 671–683.

Martín Fernández, J.A., Palarea Albaladejo, J., and Olea, R.A. (2011). Dealing with Zeros. In Compositional Data Analysis, V. Pawlowsky-Glahn, and A. Buccianti, eds. (Chichester, UK: John Wiley & Sons, Ltd), pp. 43–58.

Martín-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Mathematical Geology *35*, 253–278.

Masuda, A., Takeda, J.-I., and Ohno, K. (2016). FUS-mediated regulation of alternative RNA processing in neurons: insights from global transcriptome analysis. WIREs RNA *7*, 330–340.

Masuda, A., Takeda, J.-I., Okuno, T., Okamoto, T., Ohkawara, B., Ito, M., Ishigaki, S., Sobue, G., and Ohno, K. (2015). Position-specific binding of FUS to nascent RNA regulates mRNA length. Genes & Development *29*, 1045–1057.

McGee, W.A., Pimentel, H., Pachter, L., and Wu, J.Y. (2019). Compositional Data Analysis is necessary for simulating and analyzing RNA-Seq data. bioRxiv 564955.

Medina, I., Carbonell, J., Pulido, L., Madeira, S.C., Goetz, S., Conesa, A., Tárraga, J., Pascual-Montano, A., Nogales-Cadenas, R., Santoyo, J., et al. (2010). Babelomics: an integrative

platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Research *38*, W210–W213.

Mehta, P. (2018). Prevalence of Amyotrophic Lateral Sclerosis — United States, 2014. Morb. Mortal. Wkly. Rep. *67*, 216–218.

Mestdagh, P., Hartmann, N., Baeriswyl, L., Andreasen, D., Bernard, N., Chen, C., Cheo, D., D'Andrade, P., DeMayo, M., Dennis, L., et al. (2014). Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nature Methods *11*, 809–815.

Miller, D.J., Balaram, P., Young, N.A., and Kaas, J.H. (2014). Three counting methods agree on cell and neuron number in chimpanzee primary visual cortex. Front. Neuroanat. *8*, 69.

Mitra, K., and Lippincott Schwartz, J. (2010). Analysis of mitochondrial dynamics and functions using imaging approaches. Curr Protoc Cell Biol *46*, 4.25.1–4.25.21.

Monahan, Z.T., Rhoads, S.N., Yee, D.S., and Shewmaker, F.P. (2018). Yeast Models of Prion-Like Proteins That Cause Amyotrophic Lateral Sclerosis Reveal Pathogenic Mechanisms. Front. Mol. Neurosci. *11*, 146.

Monahan, Z., Ryan, V.H., Janke, A.M., Burke, K.A., Rhoads, S.N., Zerze, G.H., O'Meally, R., Dignon, G.L., Conicella, A.E., Zheng, W., et al. (2017). Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. Embo J. *36*, 2951–2967.

Morlando, M., Dini Modigliani, S., Torrelli, G., Rosa, A., Di Carlo, V., Caffarelli, E., and Bozzoni, I. (2012). FUS stimulates microRNA biogenesis by facilitating co-transcriptional Drosha recruitment. Embo J. *31*, 4502–4510.

Mostovich, L.A., Prudnikova, T.Y., Kondratov, A.G., Loginova, D., Vavilov, P.V., Rykova, V.I., Sidorov, S.V., Pavlova, T.V., Kashuba, V.I., Zabarovsky, E.R., et al. (2011). Integrin alpha9 (ITGA9) expression and epigenetic silencing in human breast tumors. Cell Adhesion & Migration *5*, 395–401.

Munro, S.A., Lund, S.P., Pine, P.S., Binder, H., Clevert, D.-A., Conesa, A., Dopazo, J., Fasold, M., Hochreiter, S., Hong, H., et al. (2014). Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nature Communications *5*, 5125.

Murmann, A.E., Yu, J., Opal, P., and Peter, M.E. (2018). Trinucleotide Repeat Expansion Diseases, RNAi, and Cancer. Trends Cancer *4*, 684–700.

Nakaya, T., Alexiou, P., Maragkakis, M., Chang, A., and Mourelatos, Z. (2013). FUS regulates genes coding for RNA-binding proteins in neurons by binding to their highly conserved introns. RNA *19*, 1–12.

Nana, A.L., Sidhu, M., Gaus, S.E., Hwang, J.-H.L., Li, L., Park, Y., Kim, E.-J., Pasquini, L., Allen, I.E., Rankin, K.P., et al. (2018). Neurons selectively targeted in frontotemporal dementia reveal early stage TDP-43 pathobiology. Acta Neuropathologica *137*, 27–46.

Neelagandan, N., Gonnella, G., Dang, S., Janiesch, P.C., Miller, K.K., Küchler, K., Marques, R.F., Indenbirken, D., Alawi, M., Grundhoff, A., et al. (2019). TDP-43 enhances translation of specific mRNAs linked to neurodegenerative disease. Nucleic Acids Research *47*, 341–361.

Neilsen, C.T., Goodall, G.J., and Bracken, C.P. (2012). IsomiRs – the overlooked repertoire in the dynamic microRNAome. Trends Genet. *28*, 544–549.

Neumann, M., and Mackenzie, I.R.A. (2019). Review: Neuropathology of non-tau frontotemporal lobar degeneration. Neuropathology and Applied Neurobiology *45*, 19–40.

Neumann, M., Rademakers, R., Roeber, S., Baker, M., Kretzschmar, H.A., and Mackenzie, I.R.A. (2009). A new subtype of frontotemporal lobar degeneration with FUS pathology. Brain *132*, 2922–2931.

Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark, C.M., et al. (2006). Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. Science *314*, 130–133.

Ng, A.S.L., Rademakers, R., and Miller, B.L. (2015). Frontotemporal dementia: a bridge between dementia and neuromuscular disease. Annals of the New York Academy of Sciences *1338*, 71–93.

Nolan, M., Talbot, K., and Ansorge, O. (2016). Pathogenesis of FUS-associated ALS and FTD: insights from rodent models. Acta Neuropathol Commun *4*, 617.

Noorbakhsh, J., Lang, A.H., and Mehta, P. (2013). Intrinsic noise of microRNA-regulated genes and the ceRNA hypothesis. PLoS ONE *8*, e72676.

Onesto, E., Colombrita, C., Gumina, V., Borghi, M.O., Dusi, S., Doretti, A., Fagiolari, G., Invernizzi, F., Moggio, M., Tiranti, V., et al. (2016). Gene-specific mitochondria dysfunctions in human TARDBP and C9ORF72 fibroblasts. Acta Neuropathol Commun *4*, 475.

Onyike, C.U., and Diehl-Schmid, J. (2013). The epidemiology of frontotemporal dementia. Int Rev Psychiatry *25*, 130–137.

Orang, A.V., Safaralizadeh, R., and Kazemzadeh-Bavili, M. (2014). Mechanisms of miRNA-Mediated Gene Regulation from Common Downregulation to mRNA-Specific Upregulation. Int J Genomics *2014*, 1–15.

Osella, M., Bosia, C., Corá, D., and Caselle, M. (2011). The Role of Incoherent MicroRNA-Mediated Feedforward Loops in Noise Buffering. PLoS Computational Biology *7*, e1001101.

Ou, S.-H.I., Wu, F., Harrich, D., García-Martínez, L.F., and Gaynor, R.B. (1995). Cloning and characterization of a novel cellular protein, TDP-43, that binds to human immunodeficiency virus type 1 TAR DNA sequence motifs. Journal of Virology *69*, 3584–3596.

Ozdilek, B.A., Thompson, V.F., Ahmed, N.S., White, C.I., Batey, R.T., and Schwartz, J.C. (2017). Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. Nucleic Acids Research *45*, 7984-7996.

Pachter, L. (2011). Models for transcript quantification from RNA-Seq. arXiv 1104.3889v2.

Park, Y.-Y., Kim, S.-B., Han, H.D., Sohn, B.H., Kim, J.H., Liang, J., Lu, Y., Rodriguez-Aguayo, C., Lopez-Berestein, G., Mills, G.B., et al. (2013). Tat-activating regulatory DNA-binding protein regulates glycolysis in hepatocellular carcinoma by regulating the platelet isoform of phosphofructokinase through microRNA 520. Hepatology *58*, 182–191.

Patnaik, S.K., Kannisto, E., Mallick, R., and Yendamuri, S. (2011). Overexpression of the Lung Cancer-Prognostic miR-146b MicroRNAs Has a Minimal and Negative Effect on the Malignant Phenotype of A549 Lung Cancer Cells. PLoS ONE *6*, e22379.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods *14*, 417–419.

Perry, S.W., Norman, J.P., Barbieri, J., Brown, E.B., and Gelbard, H.A. (2011). Mitochondrial membrane potential probes and the proton gradient: a practical usage guide. BioTechniques *50*, 98–115.

Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Research *29*, e45.

Phipson, B., Zappia, L., and Oshlack, A. (2017). Gene length and detection bias in single cell RNA sequencing protocols. F1000Research *6*, 595.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nature Methods *14*, 687–690.

Pinarbasi, E.S., Cağatay, T., Fung, H.Y.J., Li, Y.C., Chook, Y.M., and Thomas, P.J. (2018). Active nuclear import and passive nuclear export are the primary determinants of TDP-43 localization. Sci. Rep. *8*, 7083.

Piras, V., and Selvarajoo, K. (2015). The reduction of gene expression variability from single cells to populations follows simple statistical laws. Genomics *105*, 137–144.

Pivovarova, N.B., and Andrews, S.B. (2010). Calcium-dependent mitochondrial function and dysfunction in neurons. FEBS Journal *277*, 3622–3636.

Plewka, P., Szczesniak, M., Stepien, A., Zywicki, M., Pacak, A., Colombo, M., Makalowska, I., Ruepp, M.D., and Raczynska, K.D. (2019). FUS controls the processing of snoRNAs into smaller RNA fragments that can regulate gene expression. bioRxiv 409250.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.-C., Sun, E., Wancewicz, E., Mazur, C., et al. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nature Neuroscience *14*, 459–468.

Porta, S., Kwong, L.K., Trojanowski, J.Q., and Lee, V.M.Y. (2015). Drosha Inclusions Are New Components of Dipeptide-Repeat Protein Aggregates in FTLD-TDP and ALS C9orf72 Expansion Cases. Journal of Neuropathology and Experimental Neurology *74*, 380–387.

Postel-Vinay, S., Véron, A.S., Tirode, F., Pierron, G., Reynaud, S., Kovar, H., Oberlin, O., Lapouble, E., Ballet, S., Lucchesi, C., et al. (2012). Common variants near TARDBP and EGR2 are associated with susceptibility to Ewing sarcoma. Nature Genetics *44*, 323–327.

Pottier, C., Ravenscroft, T.A., Sanchez-Contreras, M., and Rademakers, R. (2016). Genetics of FTLD: overview and what else we can expect from genetic studies. Journal of Neurochemistry *138*, 32–53.

Prasad, A., Bharathi, V., Sivalingam, V., Girdhar, A., and Patel, B.K. (2019). Molecular Mechanisms of TDP-43 Misfolding and Pathology in Amyotrophic Lateral Sclerosis. Front. Mol. Neurosci. *12*, 1199.

Pratt, A.J., Getzoff, E.D., and Perry, J.J. (2012). Amyotrophic lateral sclerosis: update and new developments. Degenerative Neurological and Neuromuscular Disease *2012*, 1–14.

Qin, H., Lim, L.-Z., Wei, Y., and Song, J. (2014). TDP-43 N terminus encodes a novel ubiquitin-like fold and its unfolded form in equilibrium that can be shifted by binding to ssDNA. Proceedings of the National Academy of Sciences *111*, 18619–18624.

Quinn, T.P., Crowley, T.M., and Richardson, M.F. (2018a). Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics *19*, 274.

Quinn, T.P., Erb, I., Richardson, M.F., and Crowley, T.M. (2018b). Understanding sequencing data as compositions: an outlook and review. Bioinformatics *34*, 2870–2878.

Rabbitts, T.H., Forster, A., Larson, R., and Nathan, P. (1993). Fusion of the dominant negative transcription regulator CHOP with a novel gene FUS by translocation t(12;16) in malignant liposarcoma. Nature Genetics *4*, 175–180.

Rabinovici, G.D., and Miller, B.L. (2010). Frontotemporal lobar degeneration: epidemiology, pathophysiology, diagnosis and management. CNS Drugs *24*, 375–398.

Ramakers, C., Ruijter, J.M., Deprez, R.H.L., and Moorman, A.F.M. (2003). Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neuroscience Letters *339*, 62–66.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc *8*, 2281–2308.

Rascovsky, K., Salmon, D.P., Lipton, A.M., Leverenz, J.B., DeCarli, C., Jagust, W.J., Clark, C.M., Mendez, M.F., Tang-Wai, D.F., Graff-Radford, N.R., et al. (2005). Rate of progression differs in frontotemporal dementia and Alzheimer disease. Neurology *65*, 397–403.

Rascovsky, K., Hodges, J.R., Knopman, D., Mendez, M.F., Kramer, J.H., Neuhaus, J., van Swieten, J.C., Seelaar, H., Dopper, E.G.P., Onyike, C.U., et al. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain *134*, 2456–2477.

Ratti, A., and Buratti, E. (2016). Physiological Functions and Pathobiology of TDP-43 and FUS/TLS proteins. Journal of Neurochemistry *138*, 95-111.

Reber, S., Stettler, J., Filosa, G., Colombo, M., Jutzi, D., Lenzken, S.C., Schweingruber, C., Bruggmann, R., Bachi, A., Barabino, S.M., et al. (2016). Minor intron splicing is regulated by FUS and affected by ALS-associated FUS mutants. Embo J. *35*, 1504–1521.

Renton, A.E., Chiò, A., and Traynor, B.J. (2014). State of play in amyotrophic lateral sclerosis genetics. Nature Publishing Group *17*, 17–23.

Rhoads, S., Monahan, Z., Yee, D., and Shewmaker, F. (2018). The Role of Post-Translational Modifications on Prion-Like Aggregation and Liquid-Phase Separation of FUS. Int J Mol Sci *19*, 886.

Rinchetti, P., Rizzuti, M., Faravelli, I., and Corti, S. (2017). MicroRNA Metabolism and Dysregulation in Amyotrophic Lateral Sclerosis. Mol Neurobiol *55*, 2617–2630.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology *32*, 896–902.

Ritz, C., and Spiess, A.-N. (2008). qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. Bioinformatics *24*, 1549–1551.

Roberts, A., and Pachter, L. (2012). Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods *10*, 71–73.

Rogelj, B., Easton, L.E., Bogu, G.K., Stanton, L.W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N., et al. (2012). Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. Sci. Rep. *2*, 603.

Rohrer, J.D., Guerreiro, R., Vandrovcova, J., Uphill, J., Reiman, D., Beck, J., Isaacs, A.M., Authier, A., Ferrari, R., Fox, N.C., et al. (2009). The heritability and genetics of frontotemporal lobar degeneration. Neurology *73*, 1451–1456.

Romano, M., Feiguin, F., and Buratti, E. (2012). Drosophila Answers to TDP-43 Proteinopathies. Journal of Amino Acids *2012*, 1–13.

Rosen, D.R., Siddique, T., Patterson, D., Figlewicz, D.A., Sapp, P., Hentati, A., Donaldson, D., Goto, J., O'Regan, J.P., Deng, H.-X., et al. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature *362*, 59–62.

Rowland, L.P., and Shneider, N.A. (2001). Amyotrophic Lateral Sclerosis. N. Engl. J. Med. *344*, 1688–1700.

Rubie, C., Kempf, K., Hans, J., Su, T., Tilton, B., Georg, T., Brittner, B., Ludwig, B., and Schilling, M. (2005). Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. MCP *19*, 101–109.

Rudnick, P.A., Wang, X., Yan, X., Sedransk, N., and Stein, S.E. (2014). Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data. Molecular & Cellular Proteomics *13*, 1341–1351.

Ruijter, J.M., Ramakers, C., Hoogaars, W.M.H., Karlen, Y., Bakker, O., van den Hoff, M.J.B., and Moorman, A.F.M. (2009). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Research *37*, e45–e45.

Ryan, B., Joilin, G., and Williams, J.M. (2015). Plasticity-related microRNA and their potential contribution to the maintenance of long-term potentiation. Front. Mol. Neurosci. *8*, 4.

Saberi, S., Stauffer, J.E., Schulte, D.J., and Ravits, J. (2015). Neuropathology of Amyotrophic Lateral Sclerosis and Its Variants. Neurol Clin *33*, 855–876.

Saldi, T.K., Ash, P.E., Wilson, G., Gonzales, P., Lecca, A.G., Roberts, C.M., Dostal, V., Gendron, T.F., Stein, L.D., Blumenthal, T., et al. (2014). TDP-1, the Caenorhabditis elegans ortholog of TDP-43, limits the accumulation of double-stranded RNA. Embo J. *33*, 2947–2966.

Sama, R.R.K., Ward, C.L., and Bosco, D.A. (2014). Functions of FUS/TLS From DNA Repair to Stress Response: Implications for ALS. ASN Neuro *6*, 1-18.

Scekic-Zahirovic, J., Sendscheid, O., Oussini, El, H., Jambeau, M., Sun, Y., Mersmann, S., Wagner, M., Dieterle, S., Sinniger, J., Dirrig-Grosch, S., et al. (2016). Toxic gain of function from mutant FUS protein is crucial to trigger cell autonomous motor neuron loss. Embo J *35*, 1077-1097.

Schaefer, P.M., Einem, von, B., Walther, P., Calzia, E., and Arnim, von, C.A.F. (2016). Metabolic Characterization of Intact Cells Reveals Intracellular Amyloid Beta but Not Its Precursor Protein to Reduce Mitochondrial Respiration. PLoS ONE *11*, e0168157.

Schmid, B., Hruscha, A., Hogl, S., Banzhaf-Strathmann, J., Strecker, K., van der Zee, J., Teucke, M., Eimer, S., Hegermann, J., Kittelmann, M., et al. (2013). Loss of ALS-associated TDP-43 in zebrafish causes muscle degeneration, vascular dysfunction, and reduced motor neuron axon outgrowth. Proceedings of the National Academy of Sciences *110*, 4986–4991.

Schwartz, J.C., Ebmeier, C.C., Podell, E.R., Heimiller, J., Taatjes, D.J., and Cech, T.R. (2012). FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. Genes Dev. *26*, 2690–2695.

Schwartz, J.C., Wang, X., Podell, E.R., and Cech, T.R. (2013). RNA Seeds Higher-Order Assembly of FUS Protein. Cell Reports *5*, 918–925.

Seeley, W.W. (2008). Selective functional, regional, and neuronal vulnerability in frontotemporal dementia. Current Opinion in Neurology *21*, 701–707.

Seltman, R.E., and Matthews, B.R. (2012). Frontotemporal lobar degeneration: epidemiology, pathology, diagnosis and management. CNS Drugs *26*, 841–870.

Sephton, C.F., Good, S.K., Atkin, S., Dewey, C.M., Mayer, P., Herz, J., and Yu, G. (2010). TDP-43 is a developmentally regulated protein essential for early embryonic development. Journal of Biological Chemistry *285*, 6826–6834.

SEQC MAQC-III Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology *32*, 903–914.

Shan, X., Chiang, P.-M., Price, D.L., and Wong, P.C. (2010). Altered distributions of Gemini of coiled bodies and mitochondria in motor neurons of TDP-43 transgenic mice. Proceedings of the National Academy of Sciences *107*, 16325–16330.

Sheng, Z.-H. (2017). The interplay of axonal energy homeostasis and mitochondrial trafficking and anchoring. Trends in Cell Biology *27*, 403–416.

Shorter, J. (2017). Liquidizing FUS via prion-like domain phosphorylation. Embo J. *36*, 2925–2927.

Shrestha, L.B., and Heisler, E.J. (2011). Changing Demographic Profile of the United States. Congressional Research Service.

Sieben, A., van Langenhove, T., Engelborghs, S., Martin, J.-J., Boon, P., Cras, P., De Deyn, P.-P., Santens, P., Van Broeckhoven, C., and Cruts, M. (2012). The genetics and neuropathology of frontotemporal lobar degeneration. Acta Neuropathologica *124*, 353–372.

Sobue, G., Ishigaki, S., and Watanabe, H. (2018). Pathogenesis of Frontotemporal Lobar Degeneration: Insights From Loss of Function Theory and Early Involvement of the Caudate Nucleus. Front. Neurosci. *12*, 383.

Solomon, D.A., Mitchell, J.C., Konrad, M.T.S., Vance, C.A., and Mizielinska, S. (2019). Review: Modelling the pathology and behaviour of frontotemporal dementia. Neuropathology and Applied Neurobiology *45*, 58–80.

Soneson, C., Love, M.I., and Robinson, M.D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research *4*, 1521.

Sriram, G., and Birge, R.B. (2011). Emerging Roles for Crk in Human Cancer. Genes & Cancer *1*, 1132–1139.

St George-Hyslop, P., Lin, J.Q., Miyashita, A., Phillips, E.C., Qamar, S., Randle, S.J., and Wang, G. (2018). The physiological and pathological biophysics of phase separation and gelation of RNA binding proteins in amyotrophic lateral sclerosis and fronto-temporal lobar degeneration. Brain Research *1693*, 11–23.

Stauch, K.L., Purnell, P.R., and Fox, H.S. (2014). Quantitative Proteomics of Synaptic and Nonsynaptic Mitochondria: Insights for Synaptic Mitochondrial Vulnerability. J. Proteome Res. *13*, 2620–2636.

Stoica, R., De Vos, K.J., Paillusson, S., Mueller, S., Sancho, R.M., Lau, K.-F., Vizcay-Barrena, G., Lin, W.-L., Xu, Y.-F., Lewis, J., et al. (2014). ER–mitochondria associations are regulated by the VAPB–PTPIP51 interaction and are disrupted by ALS/FTD-associated TDP-43. Nature Communications *5*, 3996.

Stoica, R., Paillusson, S., Gomez Suaga, P., Mitchell, J.C., Lau, D.H., Gray, E.H., Sancho, R.M., Vizcay-Barrena, G., De Vos, K.J., Shaw, C.E., et al. (2016). ALS/FTD-associated FUS activates GSK-3β to disrupt the VAPB–PTPIP51 interaction and ER–mitochondria associations. EMBO Rep e201541726.

Stribl, C., Samara, A., Trümbach, D., Peis, R., Neumann, M., Fuchs, H., Gailus-Durner, V., Hrabě de Angelis, M., Rathkolb, B., Wolf, E., et al. (2014). Mitochondrial dysfunction and decrease in body weight of a transgenic knock-in mouse model for TDP-43. Journal of Biological Chemistry *289*, 10769–10784.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Pnas *102*, 15545–15550.

Sun, S., Ling, S.-C., Qiu, J., Albuquerque, C.P., Zhou, Y., Tokunaga, S., Li, H., Qiu, H., Bui, A., Yeo, G.W., et al. (2015). ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. Nature Communications *6*, 6171.

Talbott, S.L. (2010). Getting Over the Code Delusion. The New Atlantis *28*, 3–28.

Talbott, S.L. (2011). The Unbearable Wholeness of Beings. The New Atlantis 1–25.

Tan, A.Y., Riley, T.R., Coady, T., Bussemaker, H.J., and Manley, J.L. (2012). TLS/FUS (translocated in liposarcoma/fused in sarcoma) regulates target gene transcription via single-stranded DNA response elements. Proceedings of the National Academy of Sciences *109*, 6030–6035.

Tan, W., Pasinelli, P., and Trotti, D. (2014). Role of mitochondria in mutant SOD1 linked amyotrophic lateral sclerosis. BBA - Molecular Basis of Disease *1842*, 1295–1301.

Tank, E.M., Figueroa-Romero, C., Hinder, L.M., Bedi, K., Archbold, H.C., Li, X., Weskamp, K., Safren, N., Paez-Colasante, X., Pacut, C., et al. (2018). Abnormal RNA stability in amyotrophic lateral sclerosis. Nature Communications *9*, 744.

Tarca, A.L., Drăghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.-S., Kim, C.J., Kusanovic, J.P., and Romero, R. (2009). A novel signaling pathway impact analysis. Bioinformatics *25*, 75–82.

The Geuvadis Consortium, Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Timmons, J.A., Szkop, K.J., and Gallagher, I.J. (2015). Multiple sources of bias confound functional enrichment analysis of global -omics data. *16*, 186.

Tollervey, J.R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A.L., Župunski, V., et al. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nature Neuroscience *14*, 452–458.

Tradewell, M.L., Yu, Z., Tibshirani, M., Boulanger, M.-C., Durham, H.D., and Richard, S. (2011). Arginine methylation by PRMT1 regulates nuclear-cytoplasmic localization and toxicity of FUS/TLS harbouring ALS-linked mutations. Human Molecular Genetics *21*, 136–149.

Udagawa, T., Fujioka, Y., Tanaka, M., Honda, D., Yokoi, S., Riku, Y., Ibi, D., Nagai, T., Yamada, K., Watanabe, H., et al. (2015). FUS regulates AMPA receptor function and FTLD/ALS-associated behaviour via GluA1 mRNA stabilization. Nature Communications *6*, ncomms8098.

Umansky, S. (2018). Aging and aging-associated diseases: A microRNA-based endocrine regulation hypothesis. Aging (Albany NY) *10*, 2557–2569.

Uversky, V.N. (2017). The roles of intrinsic disorder-based liquid-liquid phase transitions in the “Dr. Jekyll–Mr. Hyde” behavior of proteins involved in amyotrophic lateral sclerosis and frontotemporal lobar degeneration. Autophagy *13*, 2115–2162.

van Blitterswijk, M., Wang, E.T., Friedman, B.A., Keagle, P.J., Lowe, P., Leclerc, A.L., van den Berg, L.H., Housman, D.E., Veldink, J.H., and Landers, J.E. (2013). Characterization of FUS Mutations in Amyotrophic Lateral Sclerosis Using RNA-Seq. PLoS ONE *8*, e60788.

van den Boogaart, K.G., and Tolosana-Delgado, R. (2013). Fundamental Concepts of Compositional Data Analysis. In Analyzing Compositional Data with R, (Berlin, Heidelberg: Springer, Berlin, Heidelberg), pp. 13–50.

Van L T Hoang, Tom, L.N., Quek, X.-C., Tan, J.-M., Payne, E.J., Lin, L.L., Sinnya, S., Raphael, A.P., Lambie, D., Frazer, I.H., et al. (2017). RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. PeerJ *5*, e3631.

Vance, C., Rogelj, B., Hortobagyi, T., De Vos, K.J., Nishimura, A.L., Sreedharan, J., Hu, X., Smith, B., Ruddy, D., Wright, P., et al. (2009). Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6. Science *323*, 1208–1211.

Vanden Broeck, L., Callaerts, P., and Dermaut, B. (2014). TDP-43-mediated neurodegeneration: towards a loss-of-function hypothesis? Trends in Molecular Medicine *20*, 66-71.

Vejnar, C.E., and Zdobnov, E.M. (2012). miRmap: Comprehensive prediction of microRNA target repression strength. Nucleic Acids Research *40*, 11673–11683.

Vera, J., Lai, X., Schmitz, U., and Wolkenhauer, O. (2012). MicroRNA-regulated networks: the perfect storm for classical molecular biology, the ideal scenario for systems biology. In MicroRNA Cancer Regulation, U. Schmitz, O. Wolkenhauer, and J. Vera, eds. (Dordrecht: Springer Netherlands), pp. 55–76.

Wang, N., Zheng, J., Chen, Z., Liu, Y., Dura, B., Kwak, M., Xavier-Ferrucio, J., Lu, Y.-C., Zhang, M., Roden, C., et al. (2019a). Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. Nature Communications *10*, 358.

Wang, T., Jiang, X., Chen, G., and Xu, J. (2015a). Interaction of amyotrophic lateral sclerosis/frontotemporal lobar degeneration–associated fused-in-sarcoma with proteins involved in metabolic and protein degradation pathways. Neurobiology of Aging *36*, 527–535.

Wang, W.-C., Lin, F.-M., Chang, W.-C., Lin, K.-Y., Huang, H.-D., and Lin, N.-S. (2009). miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinformatics *10*, 328.

Wang, W., Arakawa, H., Wang, L., Okolo, O., Siedlak, S.L., Jiang, Y., Gao, J., Xie, F., Petersen, R.B., and Wang, X. (2017). Motor-Coordinative and Cognitive Dysfunction Caused by Mutant TDP-43 Could Be Reversed by Inhibiting Its Mitochondrial Localization. Mol. Ther. *25*, 127–139.

Wang, W., Li, L., Lin, W.-L., Dickson, D.W., Petrucelli, L., Zhang, T., and Wang, X. (2013). The ALS disease-associated mutant TDP-43 impairs mitochondrial dynamics and function in motor neurons. Hum Mol Genet *22*, 4706–4719.

Wang, W., Wang, L., Lu, J., Siedlak, S.L., Fujioka, H., Liang, J., Jiang, S., Ma, X., Jiang, Z., da Rocha, E.L., et al. (2016). The inhibition of TDP-43 mitochondrial localization blocks its neuronal toxicity. Nat Med *22*, 869-878.

Wang, X., Schwartz, J.C., and Cech, T.R. (2015b). Nucleic acid-binding specificity of human FUS protein. Nucleic Acids Research *43*, 7535-7543.

Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019b). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications *10*, 380.

Ward, C.L., Boggio, K.J., Johnson, B.N., Boyd, J.B., Douthwright, S., Shaffer, S.A., Landers, J.E., Glicksman, M.A., and Bosco, D.A. (2014). A loss of FUS/TLS function leads to impaired cellular proliferation. Cell Death & Disease *5*, e1572.

White, M.A., Kim, E., Duffy, A., Adalbert, R., Phillips, B.U., Peters, O.M., Stephenson, J., Yang, S., Massenzio, F., Lin, Z., et al. (2018). TDP-43 gains function due to perturbed autoregulation in a Tardbp knock-in mouse model of ALS-FTD. Nature Publishing Group *21*, 552–563.

Wilczynska, A., and Bushell, M. (2015). The complexity of miRNA-mediated repression. Cell Death Differ *22*, 22–33.

Williams, C.R., Baccarella, A., Parrish, J.Z., and Kim, C.C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics *17*, 1013.

Winton, M.J., Igaz, L.M., Wong, M.M., Kwong, L.K., Trojanowski, J.Q., and Lee, V.M.Y. (2008). Disturbance of nuclear and cytoplasmic TAR DNA-binding protein (TDP-43) induces disease-like redistribution, sequestration, and aggregate formation. J. Biol. Chem. *283*, 13302–13309.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics *14*, 232–243.

Xiao, Z., Cheng, G., Jiao, Y., Pan, C., Li, R., Jia, D., Zhu, J., Wu, C., Zheng, M., and Jia, J. (2018). Holo-Seq: single-cell sequencing of holo-transcriptome. Genome Biology *19*, 163.

Xu, X., Nehorai, A., and Dougherty, J.D. (2013). Cell type-specific analysis of human brain transcriptome data to predict alterations in cellular composition. Systems Biomedicine *1*, 151–160.

Xu, Y.-F., Gendron, T.F., Zhang, Y.-J., Lin, W.-L., D'Alton, S., Sheng, H., Casey, M.C., Tong, J., Knight, J., Yu, X., et al. (2010). Wild-type human TDP-43 expression causes TDP-43 phosphorylation, mitochondrial aggregation, motor deficits, and early mortality in transgenic mice. Journal of Neuroscience *30*, 10851–10859.

Xu, Y.-F., Zhang, Y.-J., Lin, W.-L., Cao, X., Stetler, C., Dickson, D.W., Lewis, J., and Petrucelli, L. (2011). Expression of mutant TDP-43 induces neuronal dysfunction in transgenic mice. Mol Neurodegeneration *6*, 73.

Yanagida, M. (2009). Cellular quiescence: are controlling genes conserved? Trends in Cell Biology *19*, 705–715.

Yi, L., Pimentel, H., Bray, N.L., and Pachter, L. (2018). Gene-level differential analysis at transcript-level resolution. Genome Biology *19*, 53.

Yim, A.K.-Y., Wong, J.W.-H., Ku, Y.-S., Qin, H., Chan, T.-F., and Lam, H.-M. (2015). Using RNA-Seq Data to Evaluate Reference Genes Suitable for Gene Expression Studies in Soybean. PLoS ONE *10*, e0136343.

Young, J.J., Lavakumar, M., Tampi, D., Balachandran, S., and Tampi, R.R. (2017). Frontotemporal dementia: latest evidence and clinical implications. Therapeutic Advances in *8*, 33–48.

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology *11*, R14.

Yu, J., Cao, Q., Mehra, R., Laxman, B., Yu, J., Tomlins, S.A., Creighton, C.J., Dhanasekaran, S.M., Shen, R., Chen, G., et al. (2007). Integrative Genomics Analysis Reveals Silencing of β-Adrenergic Signaling by Polycomb in Prostate Cancer. Cancer Cell *12*, 419–431.

Yu, Y., and Reed, R. (2015). FUS functions in coupling transcription to splicing by mediating an interaction between RNAP II and U1 snRNP. Proceedings of the National Academy of Sciences *112*, 8608–8613.

Yu, Y., Chi, B., Xia, W., Gangopadhyay, J., Yamazaki, T., Winkelbauer-Hurt, M.E., Yin, S., Eliasse, Y., Adams, E., Shaw, C.E., et al. (2015). U1 snRNP is mislocalized in ALS patient fibroblasts bearing NLS mutations in FUS and is required for motor neuron outgrowth in zebrafish. Nucleic Acids Research *43*, 3208–3218.

Zhang, T., Hwang, H.-Y., Hao, H., Talbot, C., and Wang, J. (2012). Caenorhabditis elegans RNA-processing protein TDP-1 regulates protein homeostasis and life span. Journal of Biological Chemistry *287*, 8371–8382.

Zhang, T., Wu, Y.-C., Mullane, P., Ji, Y.J., Liu, H., He, L., Arora, A., Hwang, H.-Y., Alessi, A.F., Niaki, A.G., et al. (2018). FUS Regulates Activity of MicroRNA-Mediated Gene Silencing. Molecular Cell *69*, 787–801.

Zhang, Y.-J., Caulfield, T., Xu, Y.-F., Gendron, T.F., Hubbard, J., Stetler, C., Sasaguri, H., Whitelaw, E.C., Cai, S., Lee, W.C., et al. (2013a). The dual functions of the extreme N-terminus of TDP-43 in regulating its biological activity and inclusion formation. Human Molecular Genetics *22*, 3112–3122.

Zhang, Z., Almeida, S., Lu, Y., Nishimura, A.L., Peng, L., Sun, D., Wu, B., Karydas, A.M., Tartaglia, M.C., Fong, J.C., et al. (2013b). Downregulation of MicroRNA-9 in iPSC-Derived Neurons of FTD/ALS Patients with TDP-43 Mutations. PLoS ONE *8*, e76055.

Zhao, S., and Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics *16*, 97.

Zhao, W., He, X., Hoadley, K.A., Parker, J.S., Hayes, D.N., and Perou, C.M. (2014). Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics *15*, 419.

Zhou, Y., Liu, S., Liu, G., Öztürk, A., and Hicks, G.G. (2013). ALS-Associated FUS Mutations Result in Compromised FUS Alternative Splicing and Autoregulation. PLoS Genetics *9*, e1003895.

Zhu, L., Lu, Y., Xu, X.-L., and Gao, F.-B. (2012). The FTD/ALS-associated RNA-binding protein TDP-43 regulates the robustness of neuronal specification through microRNA-9a in Drosophila. Human Molecular Genetics *22*, 218–225.

Zhu, L., Xu, M., Yang, M., Yang, Y., Li, Y., Deng, J., Ruan, L., Liu, J., Du, S., Liu, X., et al. (2014). An ALS-mutant TDP-43 neurotoxic peptide adopts an anti-parallel β-structure and induces TDP-43 redistribution. Human Molecular Genetics.

Zou, Z.-Y., Zhou, Z.-R., Che, C.-H., Liu, C.-Y., He, R.-L., and Huang, H.-P. (2017). Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. Journal of Neurology, Neurosurgery & Psychiatry *88*, 540–549.

Appendix 1: Supporting Information for Chapter 2 on the role of TDP-43 and FUS in miRNA regulation

Appendix 1.1: Supplemental Figure for Chapter 2

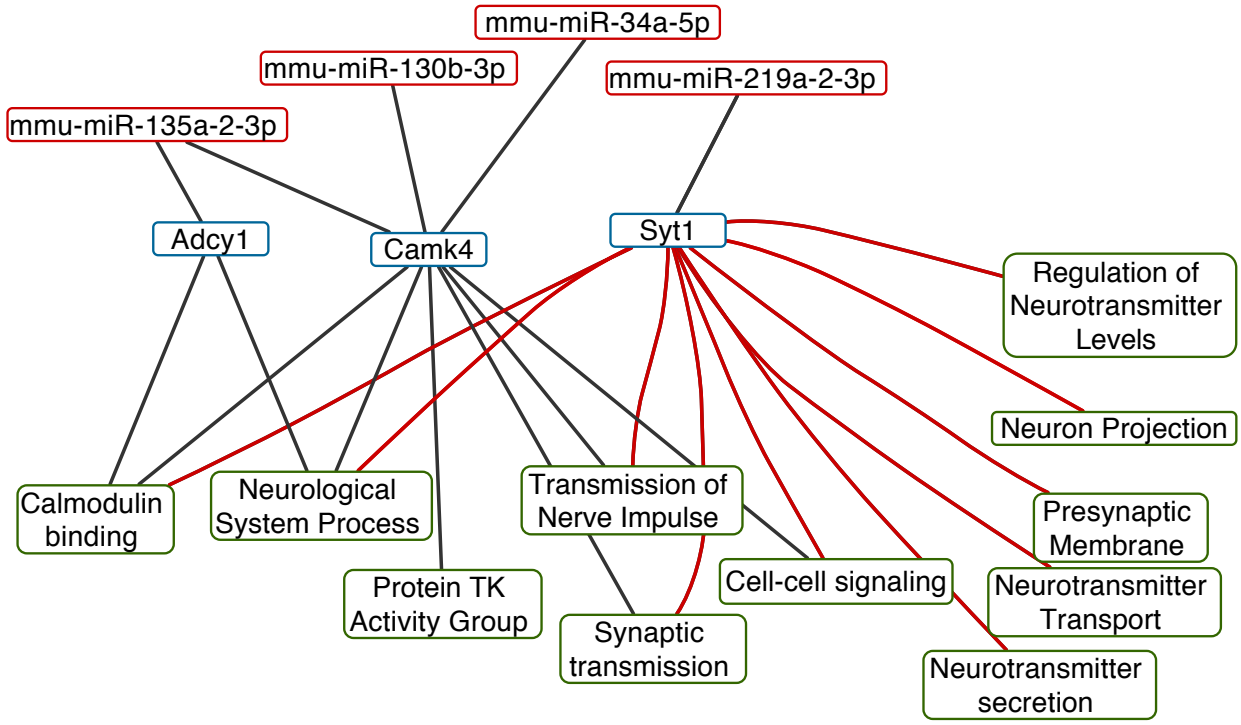

Figure A1.1: DE- and Pathway-filtered miRNA-mRNA predicted network in FUS KO Brain. *This network is the same network as shown in Figure 2.4, but in this network, only miRNAs that have a DESeq2 FDR≤0.05 and only targets that had an enriched pathway are included. This resulting network is small, with only four miRNAs and three target transcripts involved in eleven pathways. All of the miRNAs are up-regulated after FUS KO and all of the transcripts are down-regulated after FUS KO.*

Appendix 1.2: Legends for Supplemental Tables for Chapter 2

Table A1.1: a summary of the LUAD and LUSC analyses. miRNAs that are significant from the combined analysis are included, with a separate table for those that are differentially expressed and those that are not. There is a separate sheet to summarize the correlations for all miRNAs in each dataset, and a final sheet to list the TDP-43-regulated miRNAs.

Table A1.2: a list of all of the cellular processes identified by FatiGO. The first sheet contains a description of the column names for orientation, and then each sheet lists the results from the LUAD and LUSC datasets, focusing on the miRNA-mRNA interactions that are negatively correlated (down-regulated miRNA, up-regulated target; vice versa).

Table A1.3: a table to show the data to produce the hive plots shown in **Figure 2.2**. The first sheet contains a description of the columns, and then each subsequent sheet corresponds to the network in each of the four panels of **Figure 2.2**.

Table A1.4: a list of all the DESeq2 results for both miRNAs and mRNAs in the FUS KO brain dataset. Within each category, the individual transcripts/miRNAs are ordered by adjusted p-value. For the mRNA targets, there is also a column listing whether it was identified as a FUS-bound target in a previous study (Ishigaki et al., 2012).

Table A1.5: results from combining the miRNA differential expression with the Fatiscan enrichment for differentially expressed targets. The first sheet describes the columns, and the rest describe the full list of significant and non-significant miRNAs.

Appendix 2: Supporting Information for Chapter 3 on sleuth-ALR and Compositional Normalization

Appendix 2.1: *Seq datasets are compositional datasets

The term “compositional data” is defined as data where the reported values are quantitative descriptions of parts of a whole. Any data that use units relating parts to a whole (probabilities, proportions, percentages, parts per million, etc.) are thus compositional data. In other words, compositional data are restricted to reporting relative proportions of each of the components. In addition, compositional data have a "whole-sum" constraint: if one proportion increases between samples, one or more other proportions have to decrease because all of the proportions always add to the same arbitrary constant. This complicates any description of independence between components. Therefore, this kind of data requires a different statistical approach than what is usually applied to counts or continuous data, or to interval or ratio data.

Importantly, compositional data are often collected in a context where only the relative proportions matter: the absolute amounts are arbitrary with respect to the true interests of the experimenter. For example, in geology, geologists study the composition of rocks without regard to the size of the rock under study (Aitchison, 2008). Another example is an economist studying household budgets; the important aspect under study is the proportion of the budget spent on various categories like housing, food, or taxes. However, what if these disciplines wanted to know about the absolute amounts? What if the geologist wanted to know the difference in the mass of silicon present in one rock versus another? What if the economist wanted to know how much actual money was spent on taxes by one household versus another? The only way to do that is to anchor the proportions onto some value related to the whole, either the total amount

present or to some “reference value” that is unchanged between samples. Thus, the geologist would need to know the total mass of each rock, and to relate the percentages of each mineral to the total mass; the economist would need to know the total income for each household.

With *Seq experiments, the RNA molecules are not directly assayed; instead, protocols usually aliquot an equal and arbitrary amount of the total RNA from each sample to create the library for sequencing. The output for a *Seq experiment is a collection of sequenced fragments, originating from features of interest. The total number of fragments is not proportional to the total amount of RNA; instead, it is proportional to other factors, such as the experimental design, the capacity of the sequencer, and the choice of how deep the sequencing will be per sample. Because the total amount of the data (i.e. fragments) is arbitrary with respect to the value of interest (e.g. copy number changes in cell populations), *Seq experiments thus produce compositional data. The field generally already intuits this conclusion because the units used to report results from *Seq experiments are often inherently proportions: fragments per kilobase per million fragments (FPKM), transcripts per million (TPM), counts per million (CPM), etc. These units were used because it has been recognized that the sequencing depth is arbitrary with respect to the amount of RNA originally present. For example, if the same library is sequenced again at twice the depth, one could not directly compare the fragment counts between the two samples. At a minimum, one would have to account for the library size when doing any normalization.

However, we ultimately care how RNA copy numbers change between groups. To take an extreme example, if every gene was transcriptionally activated to produce twice the number of RNA molecules than what they usually produce, this would lead to an identical composition (see **Appendix 2, Table A2.3**). If an experimenter did not know about this activation, and conducted an RNA-Seq experiment, the conclusion would be that nothing had changed between

the groups, even though this is obviously not the case. Thus, *Seq experiments are restricted to only reporting how relative proportions behave unless additional assumptions are made, or additional information is collected. Compositional normalization works around these assumptions by shifting the interpretation of the data to one of pure relative comparisons: how does gene 1 behave relative to gene 2? Gene 2 may be a true "reference gene" (i.e. does not change in expression between samples), it may be a "spike-in" control, or in the absence of external information, it may reflect the average global change of RNA between samples.

Appendix 2.2: Requirements of techniques for analyzing compositional data

The following three requirements for techniques analyzing compositional data as compositional data were described by John Aitchison (Aitchison, 2003; 2008). Compositional normalization meets all three requirements, with one important exception when negative control features are available.

1. Scale invariance: analyses must treat vectors with proportional positive components as representing the same composition. In the context of *Seq experiments, this for the most part is empirically demonstrated by sequencing the same library at twice the depth. In this hypothetical example, one would expect to see twice as many fragments originating from most features in the sample with twice the sequencing depth, yet the abundance of each feature is the same. However, an important caveat for *Seq experiments, especially for single-cell RNA-Seq, is the detection limit of rare features and the dropout rate where rare features are not detected in a particular sample (Kharchenko et al., 2014; Phipson et al., 2017).

2. Subcompositional coherence: inferences about parts of a composition should be consistent, regardless of which components are included. This is important in *Seq experiments because frequently, some RNAs are selected from the total population (e.g. poly-A selection, size selection) or are depleted (e.g. rRNA depletion). Previous work has shown that correlation is subcompositionally *incoherent* and is therefore not a suitable measure for *Seq data (Lovell et al., 2015).

3. Permutation invariance: conclusions must not depend on the order of the components. Importantly, on the surface, choosing a different denominator for compositional normalization results in different estimates; however, estimates are mathematically equivalent to each other (e.g. if gene 1 increases 50% with respect to gene 2, then if gene 1 is used as the denominator instead, gene 2 will be reported as decreasing 33% with respect to gene 1, which is equivalent to the inverse of 1.5-fold increase).

What is critical for this last requirement is how one may interpret the results depending on whether negative control features are used. If they are present, using these features allows an anchoring of the data in absolute abundances that are not available using other features. However, if these features are not available, then all that is available is how some features behave relative to other features. The interpretation is restricted entirely to the relative information, and this requirement becomes imperative.

Appendix 2.3: RUVg and the Compositional Behavior of Spike-ins

Risso et al (Risso et al., 2014) had two major concerns: (1) the percentage of reads mapping to spike-ins had large systematic variation between conditions (treated versus control)

(see their Figure 4b and 4c); (2) the global behavior of spike-ins was discordant from the global behavior of genes when comparing two control libraries (see their Figure 4d and Figure 5c).

In both cases, the discrepancies arise due to how the spike-ins were added and the compositional nature of generating a library for RNA sequencing. As will be shown, observation #1 is exactly to be expected if there is a large global change in the mRNA population, and thus, it should not raise a concern. Observation #2, however, does raise a concern about how spike-ins are utilized in standard protocols. The latter concern supports an alternative method, where spike-ins are added in proportion either to the total RNA or to the number of cells, before RNA isolation. Otherwise, the spike-ins themselves are affected by the compositional changes in the data.

For observation #1, if the data are interpreted as compositional data rather than as count data, the systematic variations observed between conditions are exactly to be expected if there are any compositional shifts between them. There is reason to suspect that there are indeed is a significant global change between conditions for both datasets they examined. In the SEQC dataset (SEQC MAQC-III Consortium, 2014), sample A was Stratagene's Universal Human Reference RNA, which is a mix of RNA from ten human cancerous cell lines, one of which is a glioblastoma cell line. Sample B was Ambion's human brain reference RNA, which is a mix of RNA from the brains of 23 individuals. The two samples derive from completely different biological contexts, so one would expect a large number of changing features; indeed, the original paper identified about 5,000 genes that were differentially expressed when comparing A vs B, out of 23,437 tested (SEQC MAQC-III Consortium, 2014), representing ~20% of the tested transcriptome, roughly the same percentage as in our simulations. In the zebrafish dataset, the zebrafish were treated with gallein, which blocks Gβγ activity. The original paper identified

several histone modification enzymes whose expression was affected after gallein treatment (Ferreira et al., 2014). This suggests that there are likely global changes in histone methylation, which would result in widespread transcriptional changes. Indeed, they also detected about 5000 differentially expressed genes out of ~21,000 detected genes.

In both datasets, the percentage of reads mapped to spike-ins tends to be lower in the control group (SEQC: sample A) and higher in the treatment group (SEQC: sample B). If there is a global decrease in mRNA but the same spike-ins are added to each sample, this would increase the observed abundances (percentage) of spike-ins. This can be illustrated by a toy example (**Table A2.4**). In this example with four mRNA genes, one rRNA gene, and two spike-ins, all four mRNA genes decrease by 20%. When the RNA is sampled, the spike-ins are added, and the rRNA is depleted, the abundances of the spike-ins increase while the abundances of the mRNAs appear roughly the same. However, when the mRNAs are normalized to the spike-ins, the expected fold-changes are recovered (**Table A2.4**).

For observation #2, there are two possible explanations to explain the discrepant behavior between the spike-ins and the rest of the genes: (A) there was a dropout effect that disproportionately effects low-abundance spike-ins versus the rest of the genes; (B) there was a global change in rRNA, which would induce discrepant changes in the composition between spike-ins and mRNA genes.

It is unlikely that the observation was due to spike-in dropout. The dropout effect is related to sequencing depth: the SEQC/MAC-III observed that, even when sequencing to a billion reads, there were still new low-abundance genes being detected (SEQC MAQC-III Consortium, 2014). For a given sequencing depth, abundance, and length of a transcript, there is a probability that that transcript will not be observed. This effect is prominent in single-cell

RNA-Seq (Jia et al., 2017; Pimentel et al., 2017), but still occurs with bulk RNA-Seq. The SEQC/MAC-III consortium also previously demonstrated that low-abundance spike-ins cannot adequately recover ratios between samples (see Figure 4c in (SEQC MAQC-III Consortium, 2014)). If we reexamine Figure 4d in (Risso et al., 2014), low-abundance spike-ins are observed to have very low fold-changes, suggesting dropouts. However, if the loess local regression was repeated with the lowest quartile of spike-ins dropped, the same discrepant behavior is observed between spike-ins and genes (data not shown).

This leads to the more concerning possibility that there was a possible global change in rRNA or other non-poly-adenylated RNA. Given the variability between cells of the same type within an individual observed when conducting single-cell RNA-Seq (Piras and Selvarajoo, 2015), some variability between biological replicates is to be expected, even with rRNA. Now assume the same toy example above; in a new scenario, suppose a change between two biological replicates that led to a decrease in ribosomal RNA but no change in any mRNAs (**Table A2.5**). Then further assume that the spike-ins were accurately added in the same amount to an equal sample of total RNA, as is done in the standard protocol. Then, when the poly-A selection occurs, there is the same amount of poly-A+ mRNA by mass present in the second sample for a given amount of total RNA, but the same amount of spike-in RNA was isolated. This would then lead to a decrease in the proportion of spike-ins, and an increase in the global percentage of mRNA in the second sample. This would be a problem after normalization because the mRNAs would be observed to increase relative to the spike-ins, even though they did not change (**Table A2.5**).

This leads to the necessity of adding spike-ins relative either to total RNA isolated or to the total number of cells analyzed. The work by Lovén et al (Lovén et al., 2012) used spike-ins

proportional to the number of cells isolated, so it avoided this problem. In our work, the yeast dataset was reanalyzed using a validated reference gene, which itself was quantified relative to an external standard. A final toy example can illustrate the power of this alternative approach (**Table A2.6**). In this example, the rRNA has changed substantially, and the mRNAs are changing in different directions. One of the spike-ins is added in proportion to the total RNA before isolation, and another is added in proportion to the sample of total RNA that will be used for the library construction (an arbitrary amount). Then, after the rest of the experiment, only the spike-in added before isolation can accurately recapitulate the true fold changes (**Table A2.6**).

Current protocols specify adding spike-ins in equal amounts to a sample of total RNA after RNA isolation, but before selection and sequencing. Thus, as shown by the toy examples in **Supplemental Tables S5** and **S6**, if there are global changes in the excluded RNAs, this approach will distort the ability of the spike-ins to capture the true fold changes of the RNAs under consideration. Alleviating this concern, the thought experiments in **Table A2.4 and S6** both illustrate that spike-ins, when used carefully, can in principle capture the true fold changes.

Note that in the toy examples, the ground truth is known, and no other biases are considered. Once biases are introduced, there may be additional challenges in capturing the true fold changes of the genes under consideration. Additional work must be done to see if available spike-ins are affected differently by selection protocols or the sequencing process. However, this work must always keep in mind the compositional nature of the data being generated, especially with regard to the possible changes in excluded features.

Appendix 2.4: Extending the compositional approach to other high-throughput methods

Considering future directions for this work, it has been previously noted that compositional data is near ubiquitous in high-throughput methods available to biologists (Fernandes et al., 2014; Lovell et al., 2011; 2015). Our work focused on differential analysis in the context of RNA-Seq. One area within RNA-Seq analysis to which compositional data analysis methodology can be readily extended is in studying differential splicing. Compositional data analysis would directly be able to assess changes in the proportions of splice isoforms with respect to each other (or to the total gene expression). Compositional data analysis methodology can also be extended to other high-throughput methods. All of these other methods inherit the compositional nature described here in RNA-Seq data (Fernandes et al., 2014; Gloor et al., 2017), and all of these other methods thus have a need to consider spike-in normalization (Chen et al., 2015).

For example, consider a study comparing the binding sites of a protein across two conditions using a crosslinking immunoprecipitation sequencing method for RNA sites (CLIP-Seq) or DNA sites (ChIP-Seq). Setting aside the issues surrounding determining what constitutes a peak and quantifying those peaks, once one has estimates of peaks, this is still essentially a compositional dataset, since the total amount of fragments observed is arbitrary with respect to the total amount of RNA or DNA bound by the protein between the two conditions. Another example is metagenomics, in which researchers are interested directly in compositions of microbial communities across two populations (Gloor et al., 2017). This methodology can also be applied to proteomic or metabolomic studies, in which the normalization techniques used in those areas also assume that only a few proteins or metabolites are changing (Ejigu et al., 2013;

Rudnick et al., 2014). If one is studying a translational repressor or a stress condition where many metabolites may change dramatically, applying a compositional data analysis approach in those areas may lead to more accurate estimates.

Appendix 2.5: Handling Zeros in sleuth-ALR

A long-standing and unresolved problem in compositional data analysis is the problem of zeros (Martín Fernández et al., 2011). Zeros in either the numerator or denominator result in undefined logratios. There is also a distinction between an "essential zero", i.e. a feature with truly zero abundance, and a "rounded zero", i.e. a feature that appears to have zero abundance because the true abundance was below the limit of detection. "Rounded zeros" are considered an easier problem because they can be interpolated. In sleuth-ALR, two assumptions are made: (1) if a feature has zero counts in all samples, it is assumed to be an "essential zero" as a feature that is under transcriptional silencing, and can therefore be excluded; (2) if a feature has zero counts in some conditions but non-zero counts in others, then it is assumed to be a "rounded zero" as a feature that has the potential to be transcribed but that had low abundance because of the limit of detection related to the sequencing depth of the sample. We think that the latter is a reasonable assumption because a recent study showed that sequencing even to a depth of 1 billion reads (well beyond the typical sequencing depth in most experiments) still did not saturate the number of new, rare features being detected (SEQC MAQC-III Consortium, 2014).

Other tools (sleuth, limma, edgeR, DESeq2) have handled this issue by (A) filtering out low-abundance transcripts, and (B) introducing "pseudo-counts" (e.g. 0.5) to all features. Because of the *subcompositional coherence* principle (see **Supplemental Note 2** above), filtering is an appropriate step, and so sleuth-ALR uses the standard filtering in sleuth. However, introducing pseudo-counts can distort the composition and the dependency relationships between

features (Martín-Fernández et al., 2003). In sleuth-ALR, a multiplicative strategy that minimizes the distortion is done following (Martín-Fernández et al., 2003). Let x be a D-component composition, $x = (x_1, x_2, \ldots, x_D)$, with rounded zeros; the components are replaced in the following fashion:

$$x_i = \begin{cases} \delta_i, & \text{if } x_i = 0 \\ x_i * \left(1 - \frac{\sum \delta_i}{c_i}\right), & \text{if } x_i > 0 \end{cases}$$

where c_i is the sum constraint imposed on the data (e.g. 10^6 if TPM units are used, or the sequencing depth if counts are used). This has been shown to have the best theoretical and practical behavior of the available strategies for imputation (Martín-Fernández et al., 2003). For sleuth-ALR, $\delta_i = 0.1$ is the default for counts, and $\delta_i = 0.01$ for TPMs, though the user can choose any δ_i. These are much higher values than the much smaller value recommended by (Martín-Fernández et al., 2003), but (a) these offsets are similar to the 1 or 0.5 count offset used by previous tools, (b) they stabilize the variation among the bootstraps for low-abundance features (see **Figure A2.7**), and thus (c) they improve performance compared to recommended values for δ_i (see **Figure A2.8**). It is important to note that, within a certain range spanning at least an order of magnitude, the choice of δ_i did not significantly impact performance (see **Figure A2.8**).

Appendix 2.6: Supplemental Figures for Chapter 3

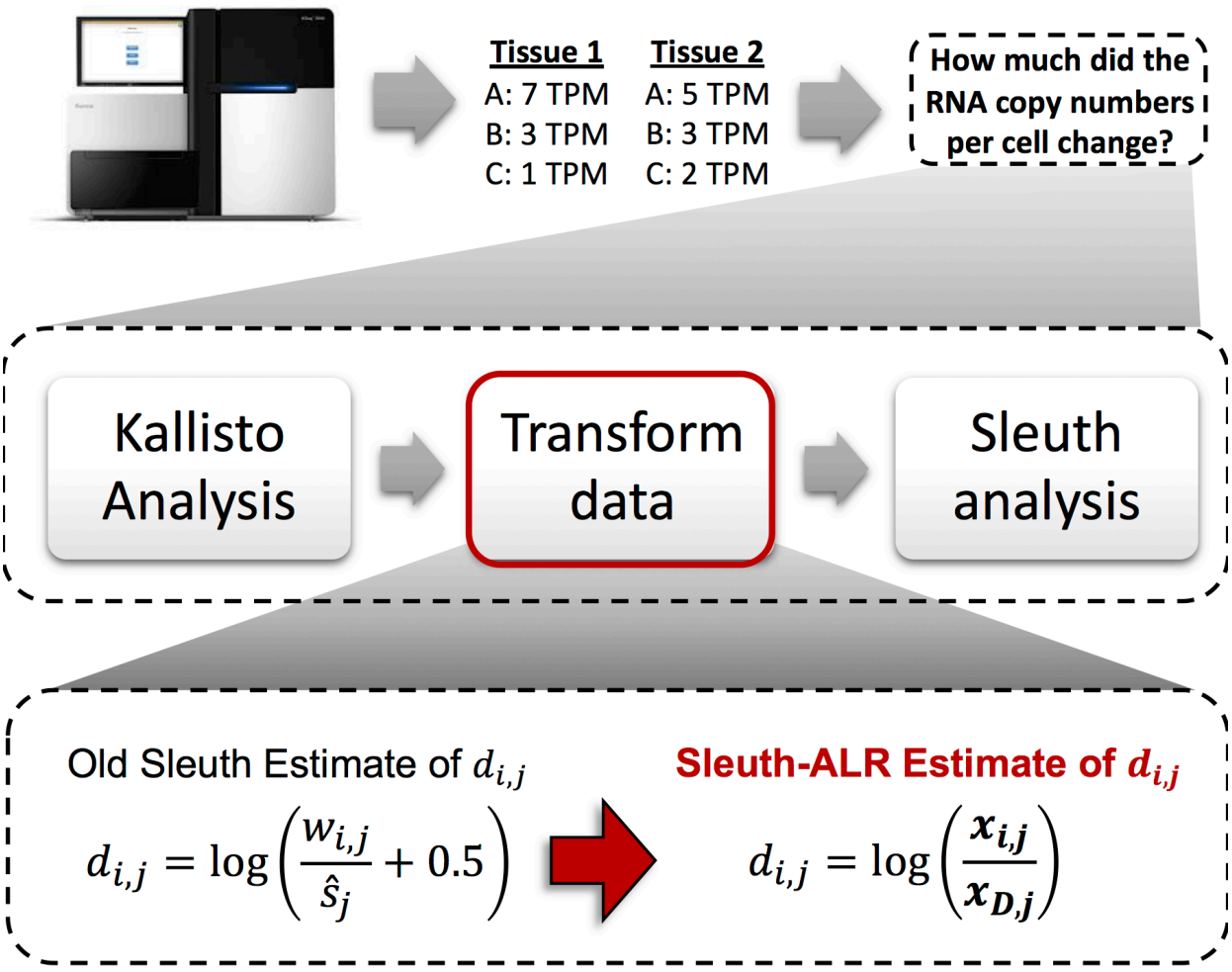

Figure A2.1: The sleuth-ALR approach for compositional normalization. *Under the sleuth model, an observation is modeled as having some error associated with it that is due to the inferential procedure. The true value is modeled as a linear combination of covariates and biological noise. In the original sleuth model (shown in the bottom left), the estimate for the noisy observation was the estimated counts for feature i in sample j, normalized by the DESeq2 size factor. This and other current normalization methods attempt to translate purely relative information to inferences about absolute changes, but only by assuming no change to the total RNA content. The proposed sleuth-ALR estimate (shown on the bottom right) is an example of how to use compositional normalization. It first focuses on abundances (TPMs) rather than estimated counts, and second normalizes the abundances by a "reference feature". This avoids having to assume only a few features change, but at the cost of not translating to inferences about absolute changes unless the chosen feature is a validated reference gene or spike-in.*

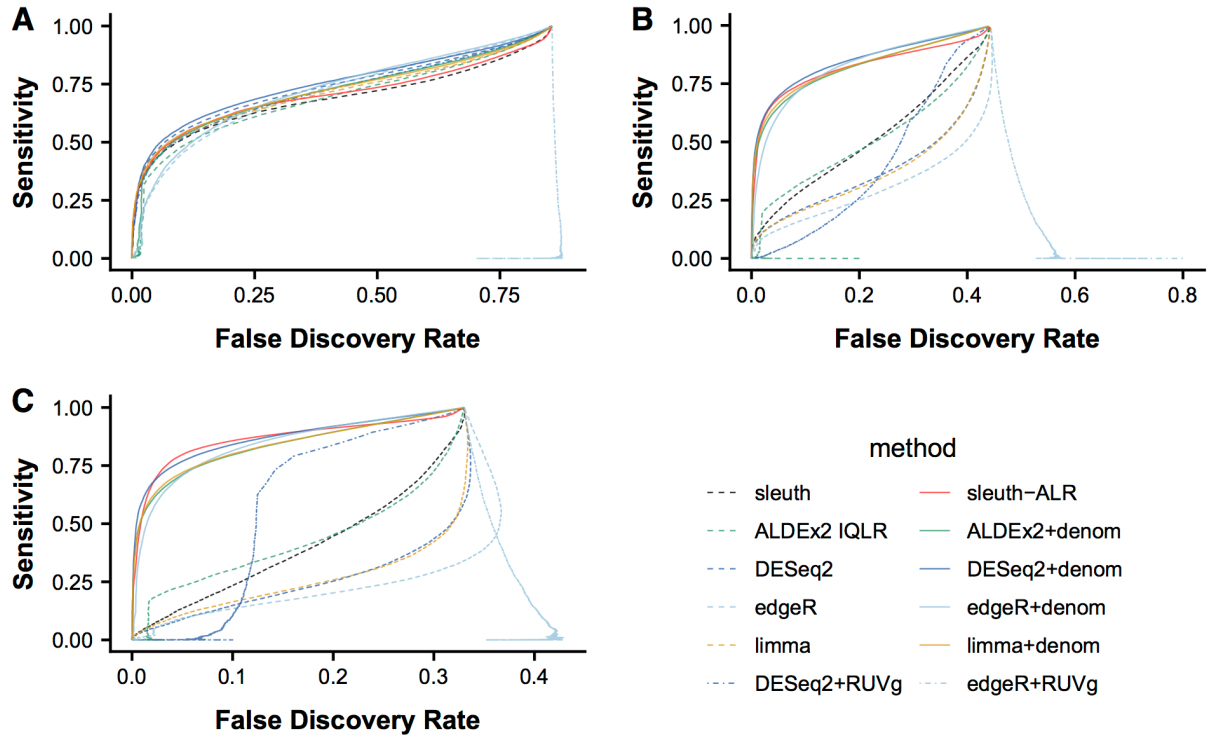

Figure A2.2: A full-range view of the simulation results, accompanying Figure 3.2. *This shows the full range of FDR and sensitivity for the three simulation studies:* ***(A)*** *"small" (5% DE; roughly equal copy numbers in each group);* ***(B)*** *"down" (20% DE; ~33% decrease in copy numbers in the experimental group); and* ***(C)*** *"up" (20% DE; ~2.8-fold increase in copy numbers in the experimental group). This shows that (1) compositional normalization has similar or superior performance throughout the full range of sensitivities and FDR, and (2) RUVg has poor performance, especially when combined with edgeR.*

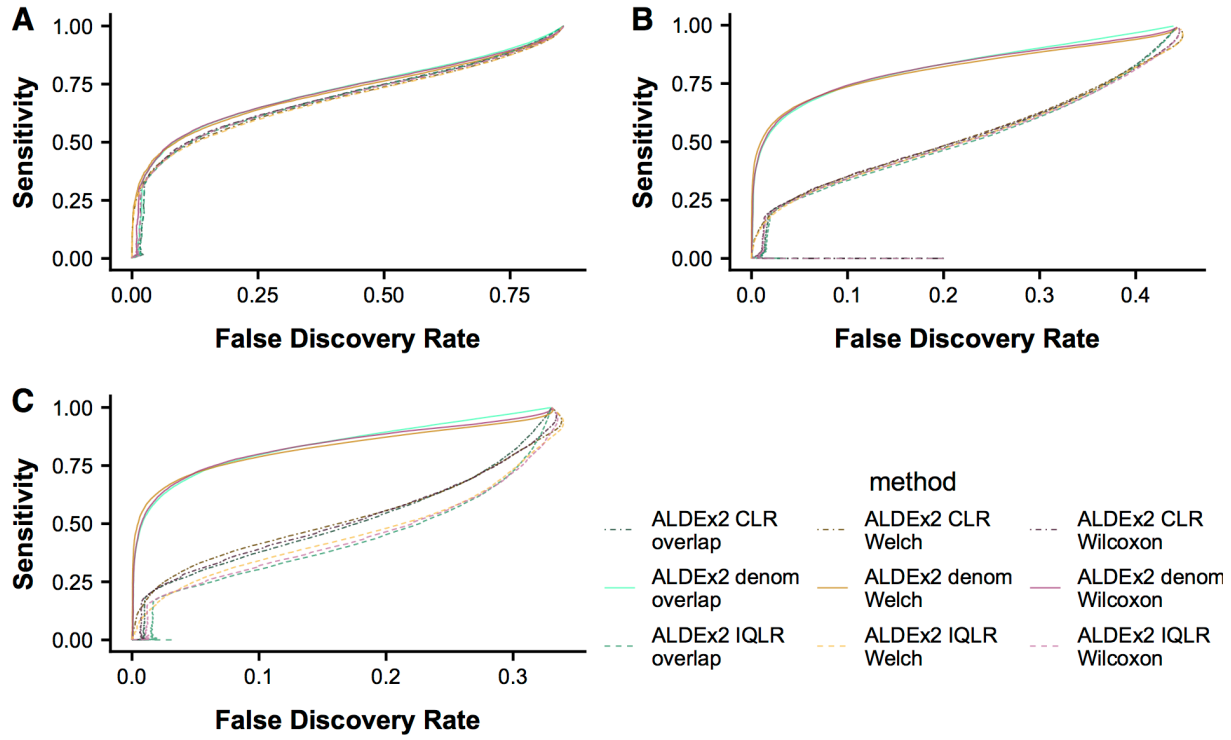

Figure A2.3: ALDEx2 performs similarly in simulations regardless of which statistical method is used*. With the same simulation studies described in* ***Figure 3.2****, the performance of ALDEx2 was compared using the Welch t-test (the recommended statistic by the developers), the non-parametric Wilcoxon test, or the reported "overlap" statistic. The overlap statistic is the posterior probability of the effect size being 0 or the opposite direction as what is reported, given the Dirichlet bootstrap samples observed. These three statistics perform approximately similarly no matter which transformation is used: CLR, IQLR, or "denom" (aka ALR, the same as used in sleuth-ALR). This remains true across all three studies:* ***(A)*** *"small" (5% DE; roughly equal copy numbers in each group);* ***(B)*** *"down" (20% DE; 33% decrease in copy numbers in the experimental group); and* ***(C)*** *"up" (20% DE; 2.8-fold increase in copy numbers in the experimental group). This is important because the Welch and Wilcoxon statistics were the ones recommended by the developers but have poor performance when there are few samples.*

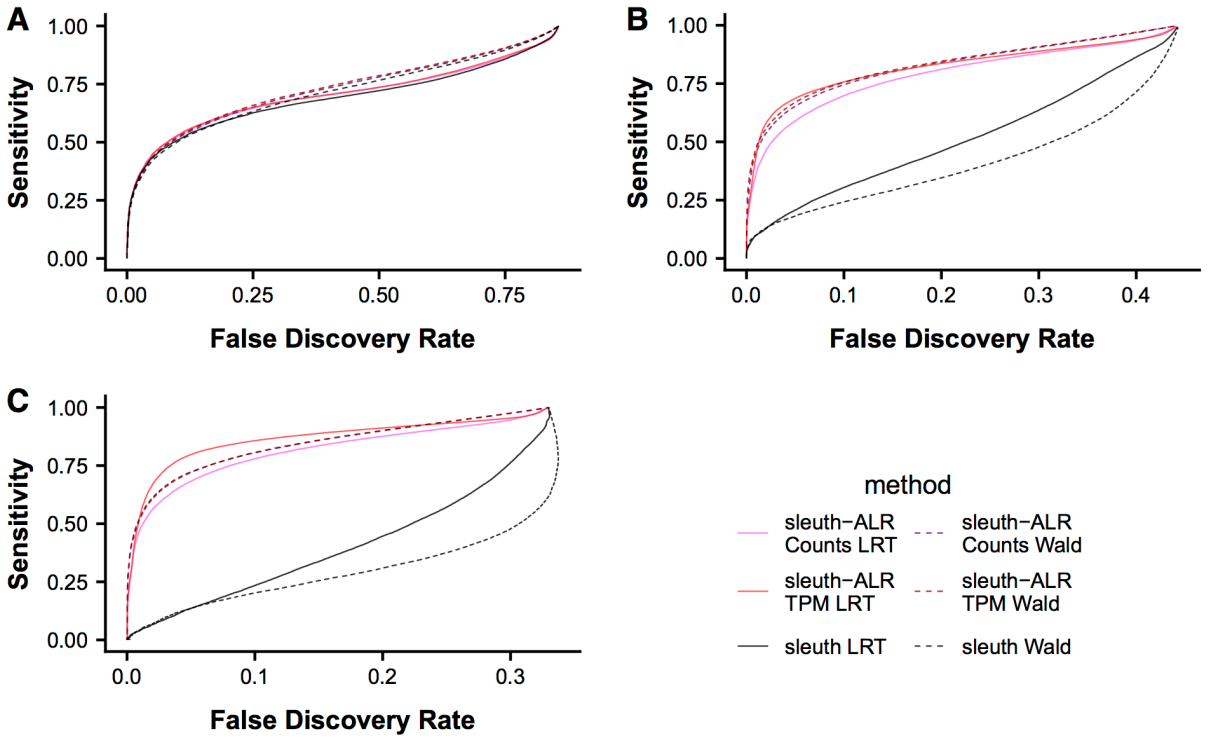

Figure A2.4: sleuth and sleuth-ALR perform similarly regardless of which statistical method or data unit is used. *With the same simulation studies described in* ***Figure 3.2****, the performance of sleuth and sleuth-ALR was compared when using the Wald test or the likelihood ratio test (LRT), as well as when using TPMs or estimated counts for modeling. All combinations perform similarly within each tool across all three studies:* ***(A)*** *"small" (5% DE; roughly equal copy numbers in each group);* ***(B)*** *"down" (20% DE; 33% decrease in copy numbers in the experimental group); and* ***(C)*** *"up" (20% DE; 2.8-fold increase in copy numbers in the experimental group).*

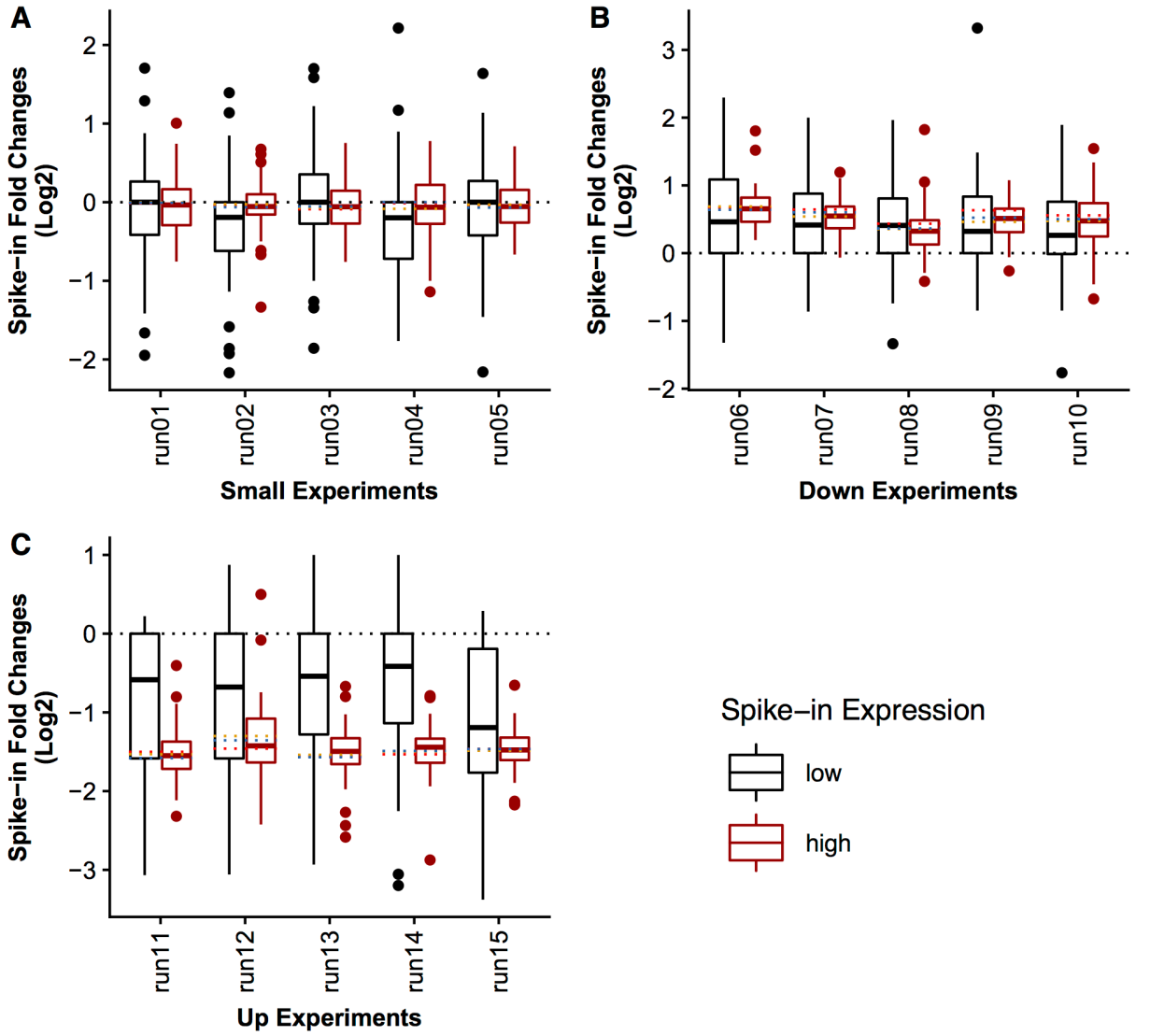

Figure A2.5: Spike-ins show a broad range of fold changes and systematic differences in studies with large shifts, accompanying Figure 3.3. *Using the "ground truth" counts from* ***polyester****, the log2 fold change was calculated for all spike-ins, and then separated by their concentration in the ERCC Mixes, with "high" expression spike-ins having an average log2 concentration of at least 3 attomoles between both mixes (N = 47 out of 92). These were the spike-ins used for normalization in* ***Figure 3.2****. Shown are boxplots of the spike-in fold changes in each experiment across the three studies:* ***(A)*** *"small" (approximately constant total RNA);* ***(B)*** *"down" (~33 decrease in total RNA); and* ***(C)*** *"up" (~2.8-fold increase in total RNA). Low-expression spike-ins tend to have a broad range of fold changes, and the high-expression spike-ins tend to have a systematic bias in fold changes in the "down" and "up" studies. For reference, the red dotted line in each run indicates the "ideal" fold change for a spike-in, if it precisely matches the reciprocal of the change in copy numbers between the control and experimental conditions; the blue and gold dotted lines indicate the fold change between conditions of the DESeq2 median-of-ratios and the sleuth-ALR geometric mean of high-expression spike-ins, respectively, suggesting that both are generally good approximations of the "ideal" fold change, and thus are good denominators for normalization.*

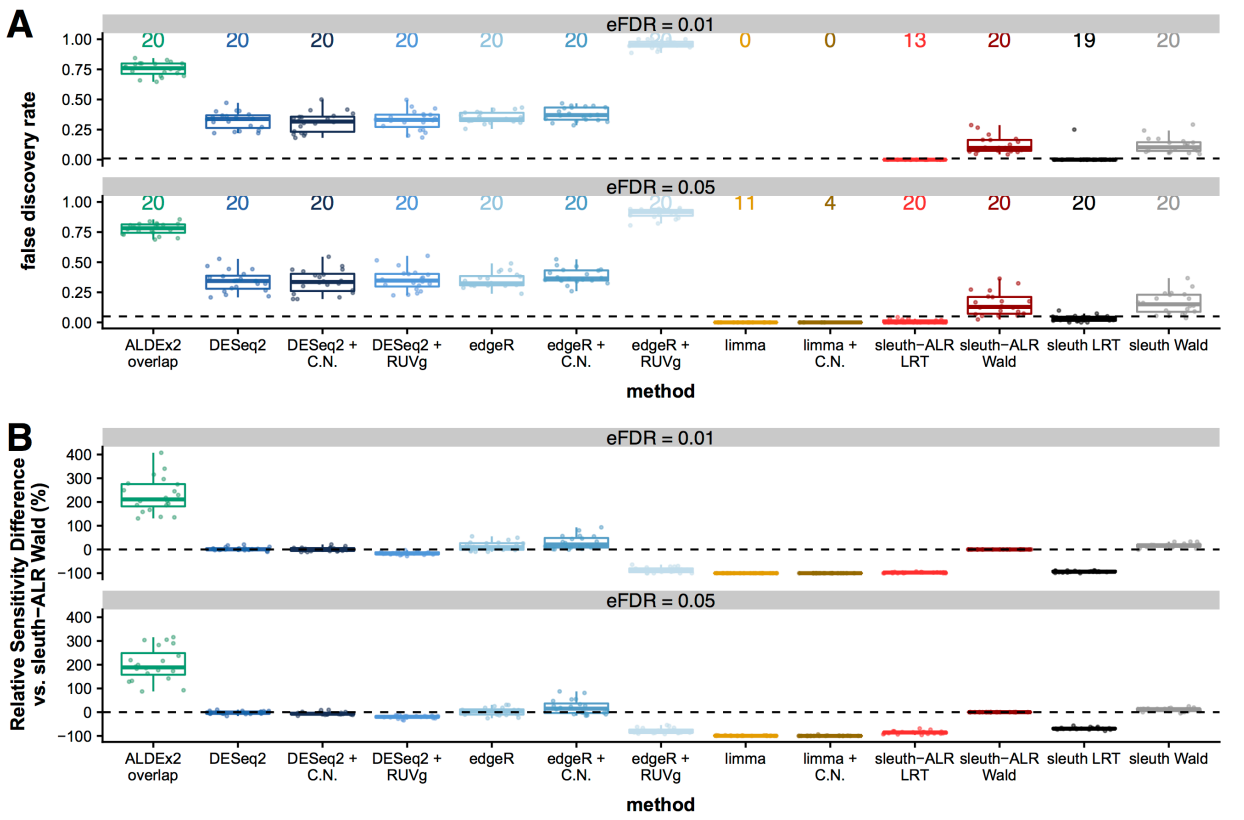

Figure A2.6: The False Discovery Rate and Relative sensitivity for the Bottomly self-consistency test at additional FDR levels. *This accompanies* ***Figure 3.4*** *in the main text. Shown here are the* ***(A)*** *False Discovery Rate, and* ***(B)*** *relative sensitivity (% change) at the FDR levels of 0.01 and 0.05.*

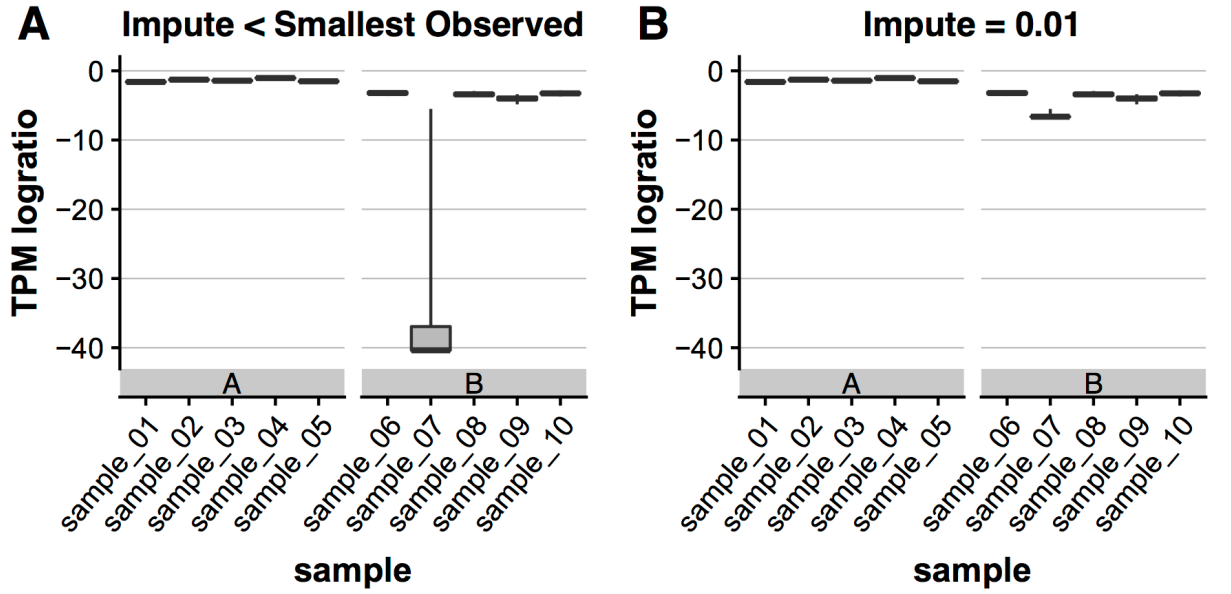

Figure A2.7: Effect of imputation value on bootstrap variation. *This depicts the summary of bootstraps variation for AAGAB-207 (ENST00000561452.5) within each sample of run #6. The true fold change for AAGAB-207 copy numbers is an 82% decrease. The recommended strategy in Compositional Data Analysis for imputing zero values is to choose a value smaller than the smallest observed value; however, because of the extremely small estimated abundances, this results in a very large variation in the bootstraps within each sample* ***(A)***. *This occurs when at least one bootstrap reports an estimated abundance of zero. Our recommendation is to follow the strategy of previous tools and choose a larger value to impute. Panel (B) shows the reduction in bootstrap variation after choosing 0.01 for the imputation. The wide variation observed in (A) resulted in a non-significant q-value (0.450), whereas the stabilized variation observed in (B) resulted in a significant q-value (0.047).*

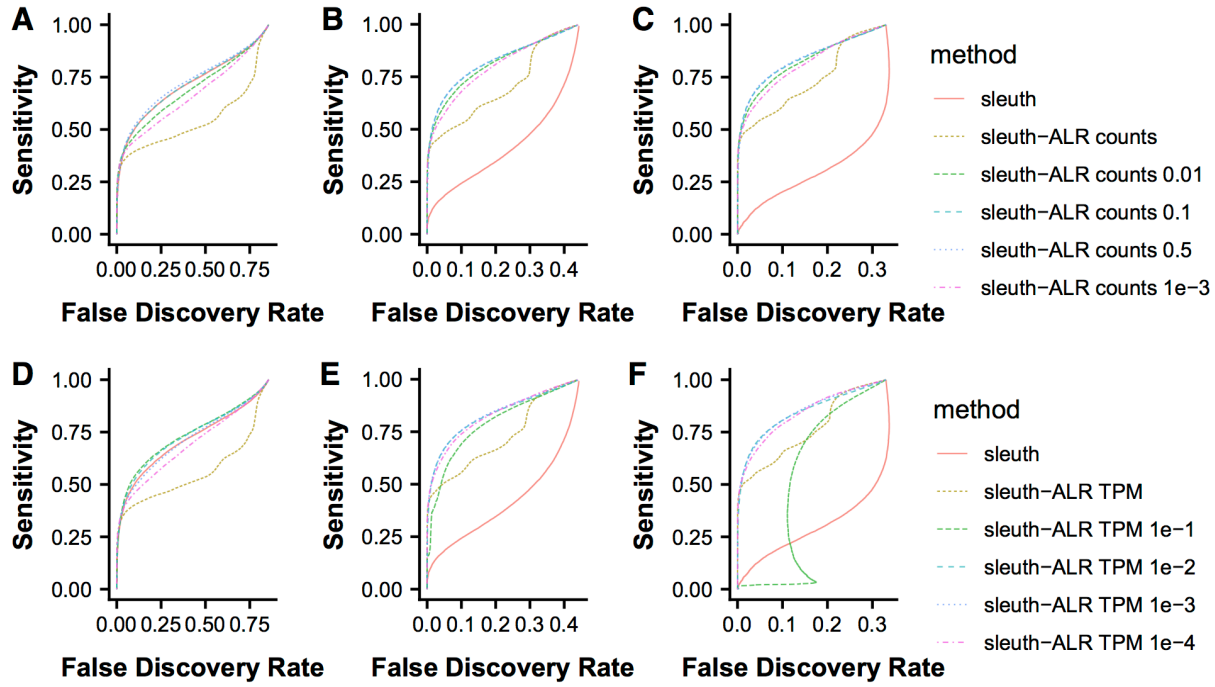

Figure A2.8: Effect of imputation on overall simulation performance*. This depicts the full sensitivity versus false discovery rate curve for different choices of imputation value, as compared to standard sleuth as well as the recommended strategy of choosing a value smaller than the smallest observed value (here depicted as "sleuth-ALR counts" for A-C and "sleuth-ALR TPM" for D-F). (**A**) and (**D**) show the results for the "small" simulation group (5% DE; <2% change in copy numbers per cell); (**B**) and (**E**) show the results for the "down" simulation group (20% DE; 33% decrease in overall copy numbers per cell); (**C**) and (**F**) show the results for the "up" simulation group (20% DE; 2.8-fold increase in overall copy numbers per cell). There is improved performance of using imputation versus no imputation, and there are only minor differences in performance in all three studies among any of the choices for imputation values except 0.1 TPM impute value, which is very high (roughly equivalent to a count imputation of 3), in the "up" study.*

Appendix 2.7: Legends for Supplemental Tables for Chapter 3

Table A2.1: Summary of Parameters for Simulation Studies. For each of the fifteen simulation runs, shown are the parameters to establish the number of differentially expressed (DE) transcripts, as well as the number that are up-regulated versus down-regulated. Only transcripts with a TPM of at least 1 were used to simulated differential expression, but the probability of differential expression was determined by the total number of transcripts (~200K). The proportion of up-regulated transcripts for the "small" study was tuned to result in similar total RNA content in both conditions. The random number generator seed was chosen solely on the basis of yielding consistency in total RNA content across each run within a study. Also shown are the actual number of DE transcripts present in the set of filtered transcripts used by all tools for each run.

Table A2.2: Total RNA Content Per Cell Per Condition for all Simulation Runs. For each of the fifteen simulation runs, shown are the total RNA copy numbers per cell for each condition. Also reported are the average change in copy numbers between the two conditions for each study.

Table A2.3: Doubling the copy numbers per cell results in the same composition. Depicted is a toy example of a simple cell with five genes of varying abundances. After an experimental manipulation, each gene has exactly double the copy numbers per cell compared to the control condition. This results in the same relative abundances, and therefore the same composition.

Table A2.4: Spike-in abundances change with large compositional shifts but still accurately capture fold changes. Depicted is another toy example using the same cell with five genes. In this case, there are large changes in the mRNA genes, but no change in the rRNA gene. Spike-ins added in equal amounts both before and after RNA isolation, show changes in their abundances, and therefore would have changes in the percentage of reads mapping to them. Despite this change in their abundance, spike-ins accurately capture the true fold changes.

Table A2.5: Spike-in abundances change discordantly when non-poly-adenylated RNA changes. Depicted is another toy example using the same cell with five genes. In this experiment, there is a small change in the rRNA, but no changes to the mRNA genes. Spike-ins were added in equal amounts after RNA isolation, to simulate the protocol used in the zebrafish dataset. Because of the unobserved rRNA change, the spike-ins are affected by the compositional change and show discordant fold changes when compared to the mRNA genes. Normalizing the mRNA genes to the spike-ins results in artefactually elevated fold changes. Thus, the discrepancy observed in the zebrafish dataset can be explained by unobserved changes in the rRNA.

Table A2.6: Spike-ins must be added before RNA isolation to accurately capture true fold changes. Depicted is a final toy example using the same cell with five genes. In this experiment, there are large and varying changes to both the rRNA and mRNA genes. Spike-ins were added in equal amounts both before and after RNA isolation. Only the spike-in added before RNA isolation can accurately capture the fold changes of the mRNA genes; the spike-in added after is itself affected by the compositional shift of the simultaneous changes in both the rRNA and mRNA genes.