Understanding Neurodegenerative Disease-Relevant Molecular Effects of Perturbagens Using a Multi-Omics Approach

by

Natasha L. Patel-Murray

B.S. Mathematical Biology, University of Michigan, 2013

Submitted to the Graduate Program of Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Systems Biology

at the

Massachusetts Institute of Technology

June 2019

Signature redacted

Author . . . . . .
Natasha Patel-Murray
Computational and Systems Biology Graduate Program
May 24th, 2019

Signature redacted

Certified by . . . . . .
Ernest Fraenkel
Professor of Biological Engineering
Thesis Supervisor

Signature redacted

Accepted by . . . . . .
Christopher B. Burge
Professor of Biology
Director, Computational and Systems Biology Graduate Program

Understanding Neurodegenerative Disease-Relevant Molecular Effects of Perturbagens Using a Multi-Omics Approach

by

Natasha L. Patel-Murray

Submitted to the Graduate Program in Computational and Systems Biology
on May 24th, 2019, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

Abstract

The complex etiology of neurodegenerative diseases is not fully understood, and the characterization of cellular pathways that are dysfunctional in these diseases is key for therapeutic development. Chemical and genetic perturbagens can probe cellular pathways to shed insight about both disease etiology and potential therapeutic targets. We analyzed the functional effects of chemical perturbagens in neurodegenerative disease models as evidenced by changes in transcriptomic, metabolomic, epigenomic, and proteomic data ("multi-omics" data). Our studies revealed novel modes of action for small molecule compounds that promote survival in a model of Huntington's Disease, a fatal neurodegenerative disorder. Integration of our multi-omics data using an interpretable network approach revealed that the autophagy and bioenergetics cellular pathways are affected by different sets of compounds that promote survival. Using staining and western blot assays, we validated the effect on autophagy for one set of compounds and found that the compounds activate this pathway. Using a cellular bioenergetics assay, we found that a second set of compounds shifts the bioenergetic flux from mitochondrial respiration to glycolysis, validating our network results. In a second study related to Huntington's Disease, we analyzed the effects of two peripheral huntingtin gene silencing techniques in mouse liver. We show that the transcriptional and metabolomic changes associated with both genetic silencing methods converge on similar cellular pathways, such as the immune response and fatty acid metabolism. As a whole, this thesis presents new insights into the functional effects of perturbagens that could impact neurodegenerative disease pathology and drug discovery.

Thesis Supervisor: Ernest Fraenkel
Title: Professor of Biological Engineering

Acknowledgments

I am incredibly grateful to have had the support of many people during my time at MIT. Thank you to my mentors, friends, and family who have been there for me over all these years.

In particular, I would like to thank my thesis advisor, Professor Ernest Fraenkel. I arrived at MIT with very little knowledge about computational biology, and I have learned a great deal from my experience in his lab. He gave me the opportunity and resources to pursue both computational and experimental research, even though I had no experience with biological experiments. I am extremely grateful for his guidance, encouragement, and patience. I would also like to thank members of my thesis committee, Professors Doug Lauffenburger and David Gifford, for their invaluable feedback and advice. A special thank you to Professor Andrew Emili for taking the time to serve as my external committee member.

The Fraenkel Lab has been a wonderful place to grow as a scientist and I am lucky to have been part of such an amazing team. Thank you to all of the members of the Fraenkel Lab, both past and present, who helped create such a warm, helpful, and caring environment: Denise MacPhail, Johnny Li, Alex Lenail, Divya Ramamoorthy, Max Gold, Michael Murphy, Aneesh Donde, Nhan Huynh, Jacquie Liu, Bryce Hwang, Brook Wassie, Anthony Soltis, Renan Escalante-Chong, Jen Wilson, Sara Gosline, Chris Ng, Nurcan Tuncbag, Xiaofeng Xin, and Ferah Yildirim. A special thanks to Mandy Kedaigle, Leila Pirhaji, Gabi Pregernig, and Tobi Ehrenberger for their tremendous advice and support. I especially would like to thank Pamela Milani and Miriam Adam for being great mentors, teachers, and friends.

I would like to thank my collaborators over the years, especially Professors Leslie Thompson and Jeff Carroll, and their lab members. I would also like to thank Sarah Kolitz for her mentorship.

Thank you to the Computational and Systems Biology program and community. A special thanks to Jacquie Carota for being welcoming and keeping everything running smoothly. I'd also like to thank the Sidney-Pacific community for being a wonderful home at MIT. Thank you to all of my friends who have made me feel at home in Boston, as well as those who provided encouragement from afar. We've had good times at soccer, foosball, zumba, birthdays, shows, dances, museums, meals, and weddings. It's been a nice balance of fun mixed in with the hard work of the PhD.

I would especially like to thank Ratul Mukerji for being incredibly caring, understanding, and supportive.

Finally, thank you to my wonderful family for their endless love and support. To my parents and brother Neil, thank you for believing in me and helping me grow into the person I am today. None of this would have been possible without you. I am forever grateful, and I would like to dedicate this thesis to you.

Contents

List of Figures and Tables

Chapter 1: Introduction

1.1 Neurodegenerative Disorders

Neurodegenerative disorders are characterized by the progressive death of specific neuronal populations. Patients typically present a mixture of clinical features, involving disruptions in behavior, cognition, and movement (Dugger and Dickson, 2017). By the presence of abnormal protein conformations, the most common disorders can be classified as amyloidoses, tauopathies, α-synucleinopathies, or TDP-43 proteinopathies (Dugger and Dickson, 2017). Many neurodegenerative disorders share the pathogenic molecular mechanism of protein misfolding and aggregation, which can cause disruptions in normal cellular processes and ultimately, neuronal dysfunction and death (Kumar et al., 2016). Impaired nucleocytoplasmic transport, mitochondrial dysfunction, and autophagy dysregulation are common features of neurodegeneration (Bhat et al., 2015; Kim and Taylor, 2017; Kiriyama and Nochi, 2015). Few treatments exist for neurodegenerative disorders, most of which target symptoms and not the underlying disease pathology (Cummings, 2017). There remains a critical need to bridge the gap between neurobiology and clinical therapy.

1.1.1 Huntington's Disease

Huntington's Disease (HD) is a rare, fatal neurodegenerative disorder affecting between 10 and 14 out of 100,000 in Western populations and is caused by a CAG triplet expansion in the huntingtin gene on chromosome 4 (McColgan and Tabrizi, 2018). This autosomal dominant mutation, discovered in 1993, encodes an expanded polyglutamine (polyQ) domain of the huntingtin protein (Kumar et al., 2015). Unaffected individuals have less than 35 CAG repeats in the polyQ domain, while the disease inflicts those with over 39 repeats with complete penetrance. Due to the instability of the polyQ expansion between generations, HD is progressive and there is genetic anticipation (Zuccato et al., 2010). The mean age of onset is 40 years, with death occurring approximately 15 to 20 years later (Ross and Tabrizi, 2011). The symptoms of HD include defects in movement, cognition, and behavioral function (Schulte and Littleton, 2011). There is no effective therapy to halt the progression of the disease (Kumar et al., 2015).

Although the exact function of huntingtin is unclear, it appears to play several roles within the cell. It is an essential protein, required for early embryonic development, and is ubiquitously expressed in most cells and within all cellular compartments (Kumar et al., 2015; Ross and Tabrizi, 2011). Huntingtin has been shown to interact with many proteins and also to be involved in transcription, antiapoptotic activity, and the trafficking processes of vesicles and organelles (Schulte and Littleton, 2011). Within brain cells, mutant huntingtin is misfolded and forms aggregates, causing loss of wild-type functions and gain of toxic new functions (Labbadia and Morimoto, 2013). These functions involve transcriptional dysregulation, impaired cytoskeletal motor functions, compromised energy metabolism, and abnormal immune activation (Labbadia and Morimoto, 2013). Together, they result in the massive striatal neuronal cell death seen in HD patients, with medium spiny neurons in the striatum selectively targeted by the disease (Ross, 2002).

1.1.2 Spinal Muscular Atrophy

Spinal Muscular Atrophy (SMA) is an autosomal recessive neurodegenerative disorder that affects motor neurons in the spinal cord and brainstem (Ahmad et al., 2016). SMA is the leading genetic cause of infant death attributed to respiratory insufficiency and is characterized by muscle weakness and severe physical disability (Farrar et al., 2017). The worldwide incidence of SMA is between 1 in 6,000 to 1 in 10,000 individuals (Ahmad et al., 2016). SMA is caused by mutations in the survival motor neuron 1 gene (*SMN1*), which result in SMN protein deficiency (Farrar et al., 2017). The full-length splicing isoform of the almost identical gene *SMN2* can partially compensate for the protein deficiency, and the copy number of *SMN2* determines the subtype of the disease, categorized by severity (Maharshi and Hasan, 2017). One FDA-approved therapy, nusinersen, exists for the treatment of SMA and works by increased the production of full-length SMN2 protein (Maharshi and Hasan, 2017).

1.1.3 Amyotrophic Lateral Sclerosis

Amyotrophic Lateral Sclerosis (ALS) is an adult onset, neurodegenerative disorder characterized by the degeneration of upper and lower motor neurons in the

brain and spinal cord (Hardiman et al., 2017). ALS has an incidence of 1-2 per 100,000 and a mean survival time of 3-5 years (Chen et al., 2018). The clinical features of ALS include progressive muscle weakness and atrophy throughout the body (Chen et al., 2018). Two FDA-approved treatments are available for ALS, but neither is effective in halting disease progression (Chia et al., 2018). The heterogeneity and unknown pathophysiology of the disease complicate diagnosis and the search for treatment options. ALS cases can be classified using differences in initial symptom presentation site, progression rate, cognitive disruption, and behavioral changes (Chen et al., 2018). Common pathophysiological features of ALS include hyperexcitability, weight loss and pain, and nearly all patients succumb to respiratory failure (Chen et al., 2018; Do-Ha et al., 2018; Hardiman et al., 2017).

Unlike Huntington's Disease, only about 10% of patients have familial disease (Chia et al., 2018). The remaining sporadic cases have unknown etiology. In 1993, the gene *SOD1* was identified to be associated with ALS, and more than 30 genes, including *C9orf72*, *TARDBP*, and *FUS*, have since been linked to the disease (Chia et al., 2018; Hardiman et al., 2017). Many of these genes are linked to protein aggregation phenotypes, which cause dysfunction in many cellular processes. Pathways known to be dysregulated in ALS include protein homeostasis, mitochondrial dynamics, RNA metabolism, cytoskeletal integrity, axonal transport, and DNA damage (Chia et al., 2018). However, mutations in the genes involved in these pathways cannot explain the majority of ALS cases, and it is unclear how environmental and lifestyle factors play a role in the disease (Hardiman et al., 2017).

1.2 Systems Biology and Omics Data Types

One approach to gain a better understanding of neurodegenerative disorder pathophysiology is to use systems biology. Since the 1990's, new experimental and bioinformatic tools have allowed for the high throughput analysis of biological data and the emergence of the modern field of systems biology (Medina, 2013; Schneider, 2013). The development of omics technologies has resulted in significant advances in our understanding of basic biology, such as the sequencing of the human genome.

Systems biology provides a computational framework to model complex biological systems, such as cells or organisms, as a whole (Altaf-Ul-Amin et al., 2014; Breitling, 2010). This approach is in contrast to the more traditional reductionist methods to investigate biology, where complex problems are broken down into a set of simpler problems (Medina, 2013). In systems biology, genome-scale measurements of the molecules that make up a system are organized in the context of molecular networks. The connections between molecules in a network can provide information about the system's structure, dynamics, design principles, and rules of control and regulation (Altaf-Ul-Amin et al., 2014; Medina, 2013).

There are many data types that can be used in systems biology, generally referred to as "omics" measurements. Omics measurements interrogate the entire set of a given level of biological molecules (Schneider, 2013). These measurements could include genome sequences, molecular structures, gene expression, binding sites and domains, protein-protein interactions, mass spectrometry, and metabolic pathways (Altaf-Ul-Amin et al., 2014). Clinically, omics data can provide valuable information about therapeutic targets and biomarkers, disease subtypes, personalized medicine, disease mechanism and drug interactions, leading to better diagnostics and new drug candidates (Kedaigle and Fraenkel, 2018; Schneider, 2013).

1.2.1 Transcriptomics

The transcriptome refers to the set of all RNA transcripts in a cell, tissue, or organism. Quantification of the RNA transcripts can provide information about gene expression, splicing events, transcriptional structure, post-transcriptional modifications and the different species of transcripts, including mRNAs, non-coding RNAs, and small RNAs (Wang et al., 2009). Currently, there are two dominant methods for measuring the transcriptome, microarrays and RNA-Seq.

Microarrays measure a set of transcript abundances via their hybridization to an array of complementary probes (Lowe et al., 2017). The transcripts are fluorescently labeled, and the fluorescence intensity at each probe indicates the transcript abundance for that probe sequence (Lowe et al., 2017). Though microarrays are cost-effective and not labor-intensive, they have limitations due to their reliance on existing knowledge

about genome sequence and probe design. Most probes do not have predictable hybridization characteristics, and the resulting microarrays have high background levels and a limited range of detection (Pozhitkov et al., 2007; Wang et al., 2009).

RNA-Seq is a method by which RNA from the entire transcriptome is extracted from cells, converted into cDNA, amplified, and then sequenced (Wang et al., 2009). There are a variety of ways to customize an RNA-Seq experiment, including different strategies for transcript enrichment, fragmentation, amplification, strand-specificity, and single or paired-end sequencing (Lowe et al., 2017). Sequencing output is the limiting factor in RNA-Seq studies because a large number of reads is required to ensure sufficient coverage of the transcriptome of interest. There are also bioinformatic challenges to working large sequencing data sets. Compared to microarrays, RNA-Seq is often the preferred technology because it has better dynamic range, requires lower input RNA amounts, is not limited by known genomic sequences, can provide information about sequence variation, and has relatively low background signal (Lowe et al., 2017; Wang et al., 2009).

1.2.2 Epigenomics

The epigenome indicates the genome-wide chromatin state. While the genome's primary sequence is relatively static across cell types, the epigenome can vary greatly and lead to distinct gene expression programs and biological functions (Kundaje et al., 2015). Histone modifications, chromatin accessibility, and DNA methylation comprise the epigenomic state, which can be assessed using several methods, such as ChIP-Seq, DNase-Seq, and ATAC-Seq (Kundaje et al., 2015).

Histone modifications contribute to the dynamic nature of chromatin. The combination of modifications, such as acetylation and methylation, across the genome can be read as a "histone code" (Jenuwein, 2001). Histone modifications can provide information about transcriptional states and whether a region is euchromatic or heterochromatic (Jenuwein, 2001). Acetylation is typically associated with accessible chromatin, while methylation can be associated with open or compacted chromatin.

For example, the H3K4me1, H3K4me3 and H3K36me3 marks are all associated with open and transcribed chromatin, while the H3K9me3 and H3K27me3 are

associated with compacted and repressed chromatin (O'Geen et al., 2011). Also, each mark has distinct characteristics and targets. H3K4me1 is associated with transcriptional enhancers, H3K4me3 is associated with gene promoter regions, and H3K36me3 is associated with transcribed regions of the genome (Heintzman et al., 2007; O'Geen et al., 2011). Though H3K9me3 and H3K27me3 are both associated with repressive chromatin, each mark is associated with distinct sets of target genes (O'Geen et al., 2011).

ChIP-Seq is a technique for assaying protein-DNA binding and can be used to identify genome-wide profiles of transcription factors, histone modifications, DNA methylation and nucleosome positioning (O'Geen et al., 2011; Park, 2009). ChIP-Seq uses antibodies to select specific proteins or nucleosomes, which are bound to DNA fragments. The DNA fragments of interest are sequenced directly and can be used to identify regions of the genome bound to the protein or nucleosome (O'Geen et al., 2011). ChIP-Seq is often limited by cost, sequencing depth, input material, and antibody quality (Park, 2009).

Unlike ChIP-Seq, DNase-Seq is a method of mapping DNase I hypersensitive sites, which can be used to predict the location of genetic regulatory elements (Boyle et al., 2008). DNase-Seq uses the DNase I nuclease to digest chromatin. DNase I preferentially cuts at a DNase I hypersensitive sites and inserts a linker that can be identified by sequencing (Boyle et al., 2008). Sequencing reads can be used to identify DNase I hypersensitivity sites, which are associated with regulatory regions such as enhancers, promoters, silencers, insulators, and locus control regions (Boyle et al., 2008). This method can reveal novel relationships between chromatin accessibility, transcription, and transcription factor occupancy (Thurman et al., 2012).

ATAC-Seq is another sequencing-based assay that can be used to investigate chromatin accessibility. Unlike DNase-Seq, ATAC-Seq uses the Tn5 transposase to insert sequencing adaptors into accessible regions of chromatin. Sequencing reads can then be used to identify open-chromatin regions across the epigenome (Buenrostro et al., 2013). ATAC-Seq can simultaneously interrogate transcription factor occupancy, nucleosome positions in regulatory sites, and chromatin accessibility genome-wide.

Compared to DNase-Seq, it also requires less input material, is faster, and is less labor-intensive (Buenrostro et al., 2013).

1.2.3 Metabolomics

The set of all metabolites in a cell, tissue, or organism makes up the metabolome (Clish, 2015). Metabolomics involves the measurement of endogenous and exogenous small-molecule compounds that are the substrates and products of biochemical reactions (Liu and Locasale, 2017). Metabolomics platforms include mass spectrometry-based methods and nuclear magnetic resonance (Liu and Locasale, 2017). There are several protocols for mass spectrometry, and the choice of protocol depends on the type of molecules of interest. For example, lipids and polar metabolites have different sample preparation and chromatography protocols (Clish, 2015). Mass spectrometry methods can also be classified as targeted or untargeted. Targeted mass spectrometry refers to the measurement of absolute concentrations of molecules and typically requires reference standards (Liu and Locasale, 2017). As a result, relatively few metabolites can be measured in this way. Untargeted mass spectrometry, on the other hand, profiles thousands of unknown features. However, metabolite identification in untargeted mass spectrometry presents a significant bottleneck in deriving biological knowledge from such studies (Dunn et al., 2013). Other limitations for metabolomics include batch effects, instrument variation, and the diversity of technologies without standard operating procedures (Clish, 2015; Liu and Locasale, 2017).

1.2.4 Proteomics

The proteome refers to the set of all proteins present in a cell, tissue or organism, and is complex, dynamic, and represents the functional information of genes (Aslam et al., 2017). The characterization of the proteome includes expression, structure, functions, interactions and post-translational modifications, such as phosphorylation or ubiquitination (Aslam et al., 2017). Mass spectrometry is a key technology in quantifying protein expression. There are different mass spectrometry protocols, such as tandem mass spectrometry or MS3 (Aslam et al., 2017; McAlister et al., 2014). In bottom-up proteomics, the proteins in a sample are first broken down into peptides, the mass

spectrometry analysis is performed on the individual peptides, and the information is combined together to reveal protein identities (Gundry et al., 2009). There are many customizable steps to the mass spectrometry protocols, including protein purification and digestion, peptide tagging, peptide enrichment, and peptide cleanup (Gundry et al., 2009). The examination of the proteome is limited by cost, access to facilities with skilled personnel, and databases for protein mapping (Aslam et al., 2017).

1.3 Computational Modeling of Biological Systems

To interpret the wealth of data generated by the omics technologies, computational techniques must be implemented. Many methods have been developed to reduce the dimensionality of large data sets, cluster and classify samples or features based on molecular profiles, correlate sets of molecules or phenotypes, find biologically meaningful pathway enrichments, and understand the connectivity and dependencies of molecules (Altaf-Ul-Amin et al., 2014; Kedaigle and Fraenkel, 2018; Parikshak et al., 2015; Prathipati and Mizuguchi, 2015; Wood et al., 2015). These methods can help define the functions of unknown omics molecules, predict and detect interacting proteins or complexes, analyze evolution of sequences or molecules across species, integrate information across omics data types, determine the most important molecules within the omics data, find biomarkers for disease diagnosis, identify drug targets and drug interactions, and compare different biological mechanisms (Altaf-Ul-Amin et al., 2014).

Data-driven methods for analyzing multi-dimensional data include dimensionality reduction and correlation approaches (Wood et al., 2015). Dimensionality reduction serves to project the multi-dimensional data onto a smaller number of dimensions. It can reduce the complexity of the data, increase interpretability, and be used to classify samples. Commonly used techniques include principal component analysis and t-SNE (Giuliani, 2017; Oliveira et al., 2018). Other correlation approaches, such as hierarchical clustering and partial least squares regression, can assess the dependencies between features or group samples and features based on their similarity (Bourgeois and Kreeger, 2017; MacLachlan et al., 2017; Si et al., 2014).

Pathway analysis extends correlation approaches by using prior knowledge about gene coregulation (Wood et al., 2015). Databases such as the Kyoto

Encyclopedia of Genes and Genomes (KEGG), the Gene Ontology (GO), Reactome, and WikiPathways relate genes and other molecules based on their mechanistic connections or functional relationships (Ashburner et al., 2000; Carbon et al., 2019; Fabregat et al., 2018; Kanehisa et al., 2017; Slenter et al., 2018). Enrichment tools, such as GOrilla and IMPaLA, use these databases to perform statistical enrichment for given gene or metabolite sets (Eden et al., 2009; Kamburov et al., 2011).

A more sophisticated technique to model omics data is network analysis. Gene regulatory networks, protein interaction networks, and Bayesian networks can be leveraged to identify connections between molecules and visualize the dependencies in the omics data (Wood et al., 2015). Given *a priori* information, causality can even be inferred from undirected networks. There are a variety of computational tools that use interaction networks to integrate different omics data. These tools include ANIMA, NetworKIN, MAGNETIC, PARADIGM, PathLinker, HotNet2, Hierarchical HotNet, PIUMet and Omics Integrator, and the choice of tool depends on the types of input data, background interactome, and biological questions of interest (Deffur et al., 2018; Leiserson et al., 2015; Linding et al., 2007; Pirhaji et al., 2016; Reyna et al., 2018; Ritz et al., 2016; Tuncbag et al., 2016; Vaske et al., 2010; Webber et al., 2018). The resulting networks constructed by each method are graphs with nodes that represent genes, proteins or metabolites, and edges that represent the potential links between them, weighted with interaction probabilities. Using machine learning algorithms, such as the Prize Collecting Steiner Forest algorithm, subnetworks most relevant to the omics data can prioritize biological pathways for further interrogation (Tuncbag et al., 2013).

Depending on the available data and prior knowledge, other types of models can be implemented. When several parameters are known *a priori*, mathematical models can be built using differential equations to provide detailed predictions of a system's dynamics (Simeoni et al., 2018). However, the information regarding parameters is generally unknown and there are usually too many variables to model. In this case, logic models can be used to predict the behavior of a system. Logic models can provide a good approximation of a system without the need for a large parameter space (Wynn et al., 2012).

1.4 Perturbagens in Biological Systems

To understand the link between particular biological pathways and the dynamics of a system, perturbagens can be applied. These perturbagens refer to any condition that can alter cellular state and can be broadly classified as chemical or genetic (Keenan et al., 2018). Chemical perturbagens are small molecule compounds. These include FDA-approved drugs and research toolkit compounds. Genetic perturbagens include gene silencing, editing or overexpression constructs, such as antisense oligonucleotides, RNAi, and CRISPR/Cas constructs (Lamb et al., 2006).

Previous large-scale studies have focused on profiling perturbagens in different cell models. For instance, a gene regulatory network was developed using transcriptional signatures from 1,484 yeast gene deletion mutants (Kemmeren et al., 2014). The analysis of these genetic perturbagens led to the identification of several gene-specific repressors (Kemmeren et al., 2014). The Connectivity Map consortium has created a catalog of thousands of reduced-representation gene expression profiles from chemical and genetic perturbagens in multiple human cell types (Lamb et al., 2006; Subramanian et al., 2017). The NIH Library of Integrated Network-Based Cellular Signatures (LINCS) program is a collaborative effort to create a network-based understanding of human biology by cataloging gene expression, proteomic, cell morphology, and epigenomic profiles from tens of thousands of pharmacological, genetic, and environmental perturbagens in multiple cell lines (Keenan et al., 2018; Litichevskiy et al., 2018).

The response of a system to perturbagens can reveal valuable information about the mechanisms of the system. In a disease context, the functional pathways altered by perturbagens that ameliorate a disease phenotype can be studied to reveal new therapeutic targets and aid in drug discovery (Wood et al., 2015). Omics technologies and the accompanying computational methods can reveal the functional pathways affected by the perturbagens. For chemical perturbagens, these functional pathways represent the modes of action of the compounds (Mulas et al., 2017).

1.4.1 Chemical Perturbagens in Drug Discovery

Traditionally in drug discovery, chemical perturbagens with potentially unknown targets are screened against a disease phenotype, with the goal of finding a therapeutic candidate. These high-throughput screens can quickly find lead compounds for drug development, but typically require thousands of molecules to be tested. They also often require serendipity, provide little mechanistic insight, and are unable to capture complex phenotypes (Varma et al., 2008). In the context of neurodegenerative disorders, screening assays are typically based on aggregation or cell death phenotypes, which only represent a subset of the multifaceted pathophysiology of these diseases (Varma et al., 2008).

Once a small molecule compound is identified to be disease-modifying, it must undergo vigorous testing (Mohs and Greig, 2017). This follow-up is crucial to understand the compounds' effects at the cell, tissue, and organism levels. Compounds' known binding targets, or mechanisms of action, do not necessarily dictate their complex downstream functional effects, or modes of action (MoAs) (Tulloch et al., 2018). Unpredicted MoAs of drug candidates can result in clinical trial failure, an expensive outcome of 86% of drug development programs (Wong et al., 2018).

Bioactivity and drug response data for chemical perturbagens have been recorded in several online databases. The Genomics of Drug Sensitivity in Cancer database is the largest public resource for drug sensitivity and response information in hundreds of cancer cell lines (Yang et al., 2013). DrugBank contains information about drug targets and interactions, as well as the influence of hundreds of drugs on metabolite, gene expression, and protein expression levels (Wishart et al., 2018). ChEMBL is a bioactivity database with compound targets, structures, and phenotypic data (Gaulton et al., 2017). Similarly, the Drug Repurposing Hub is a collection of detailed annotations for compounds that have been tested in human clinical trials or are marketed around the world (Corsello et al., 2017). These databases can be used to compare perturbagens or identify perturbagens with a desired effect for repurposing or synergy studies.

1.4.2 The Search for Modes of Action

The search for compounds' MoAs is a challenge that can stretch across decades (Mohs and Greig, 2017). Previous studies have shown that in some cases, MoAs can be inferred from omics data under restricted conditions, such as when a reference compound with a known MoA was available. For example, transcriptomic signatures of small molecule compounds were correlated with their sensitivity patterns across human cancer cell lines to identify their MoAs (Rees et al., 2016). Using reference compounds with known MoAs, compounds with similar profiles to the reference were found to have shared MoAs. In a study to combat antibiotic resistance, regression analysis was applied to untargeted metabolomic data to predict the MoAs of uncharacterized antimicrobial compounds in bacteria (Zampieri et al., 2018). MoAs could only be predicted for the uncharacterized compounds by their metabolite profile similarity to the reference compounds. Because of their reliance on reference compounds, both of these methods require additional data to predict novel MoAs.

As mentioned previously, the Connectivity Map and LINCS projects have created catalogs of gene expression and proteomics profiles from pharmacological perturbagens for the purpose of determining MoAs. The Connectivity Map includes gene expression data for 978 landmark genes from their L1000 assay (Lamb et al., 2006). The expression of an additional 11,350 genes is claimed to be inferred from the landmark genes (Lamb et al., 2006; Subramanian et al., 2017). The LINCS project extended the Connectivity Map to include proteomic data for epigenetic histone modifications and 96 phosphoproteins using the global chromatin profiling and P100 assays, respectively (Litichevskiy et al., 2018). Connectivity scores were calculated for each compound pair to highlight the similarities between compounds with similar gene expression or proteomic profiles. The catalog and connectivity approach allow for comparisons between small molecules, but they do not provide direct knowledge about unknown modes of action unless reference compounds are compared. The relatively small set of genes, proteins, and phosphosites that are directly measured limit the feature space and could be problematic when compounds affect unmeasured entities. Because compounds can have unpredicted effects, the inferred expression for the extra 11,350 genes could be incorrect. Also, the connectivity approach does not reveal the

connections between the genes or proteins that lead to functional alterations in biological pathways.

Approaches that require reference compounds must be improved to permit discovery of novel MoAs. To overcome the reference compound dependency, a regulatory network algorithm, DeMAND, was developed to use gene expression profiles to identify compounds' protein targets and activity modulators (Woo et al., 2015). This tool was able to identify novel MoAs for compounds in B cells and breast cancer cells, and it did not require reference compounds. However, though individual protein targets were predicted to be relevant to a compound's pharmacological effect, the tool did require prior knowledge of context-specific gene-regulatory interactions. This limits its general use because in many disease contexts, especially neurodegenerative disorders, such interactions are not fully characterized. Elucidating the MoAs of compounds remains a major hurdle and novel approaches with general application will be crucial to increasing the success rate of clinical trials and drug repurposing efforts (Tulloch et al., 2018; Wehling, 2009).

1.5 Overview of Thesis Contents

The goal of this thesis is to understand the effects of perturbagens in models of neurodegenerative disorders using multiple omics data and computational analyses. In Chapter 2, we describe a general multi-omics network approach for identifying modes of action (MoAs) of chemical perturbagens. We sought to identify novel MoAs for compounds identified in the search for drugs to treat Huntington's Disease. We gathered transcriptomic, metabolomic, epigenomic, and proteomic data from HD cells treated with a subset of these compounds. To find the underlying MoAs for each group, we used a feature selection approach that leverages prior biological data encoded in a molecular interaction network. A machine-learning network optimization algorithm applied to this large interactome reveals the altered biological pathways. Finally, we experimentally validated the most HD-relevant MoAs.

In the context of HD, Appendix A describes work done in collaboration with Dr. Jeff Carroll at Western Washington University. We compared the transcriptomic and metabolomic effects of two huntingtin gene silencing techniques in the mouse liver. We

found significant transcriptomic changes, but few metabolomic changes. Similar cellular pathways were affected by both silencing techniques.

Chapter 3 and Appendix B describe work done in collaboration with the NeuroLINCS consortium. The NeuroLINCS consortium explores the molecular mechanisms underlying SMA and ALS. In Chapter 3, we describe a cell freezing protocol to perform ATAC-Seq on motor neurons derived from induced pluripotent stem cells derived from patients with SMA. The loss of *SMN1* in these cells can be considered a genetic perturbagen. Flash frozen cells had disrupted chromatin, whereas cryopreserved cells retained their chromatin structure for ATAC-Seq profiling. In Appendix B, we characterize motor neurons from induced pluripotent stem cells derived from patients with ALS. These cells carry hexanucleotide expansions in *C9orf72*. Network analysis of the changes in omics data induced by this genetic perturbagen revealed causal and compensatory cellular pathways in ALS.

1.6 References

Ahmad, S., Bhatia, K., Kannan, A., and Gangwani, L. (2016). Molecular Mechanisms of Neurodegeneration in Spinal Muscular Atrophy. J. Exp. Neurosci. *10*, 39–49.

Altaf-Ul-Amin, M., Afendi, F.M., Kiboi, S.K., and Kanaya, S. (2014). Systems Biology in the Context of Big Data and Networks. Biomed Res. Int. 1–11.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Aslam, B., Basit, M., Nisar, M.A., Khurshid, M., and Rasool, M.H. (2017). Proteomics: Technologies and Their Applications. J. Chromatogr. Sci. *55*, 182–196.

Bhat, A.H., Dar, K.B., Anees, S., Zargar, M.A., Masood, A., Sofi, M.A., and Ganie, S.A. (2015). Oxidative stress, mitochondrial dysfunction and neurodegenerative diseases; a mechanistic insight. Biomed. Pharmacother. *74*, 101–110.

Bourgeois, D., and Kreeger, P. (2017). Partial Least Squares Regression Models for the Analysis of Kinase Signaling. Methods Mol. Biol. *1636*, 523–533.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell *132*, 311–322.

Breitling, R. (2010). What is systems biology? Front. Physiol. *1*, 9.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N.L., Lewis, S.E., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., et al. (2019). The Gene Ontology Resource:

20 years and still GOing strong. Nucleic Acids Res. *47*, D330-8.
Chen, H., Kankel, M.W., Su, S.C., Han, S.W.S., and Ofengeim, D. (2018). Exploring the genetics and non-cell autonomous mechanisms underlying ALS/FTLD. Cell Death Differ. *25*, 646–660.
Chia, R., Chiò, A., and Traynor, B.J. (2018). Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. *17*, 94–102.
Clish, C.B. (2015). Metabolomics: an emerging but powerful tool for precision medicine. Mol. Case Stud. *1*, a000588.
Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M., et al. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. Nat. Med. *23*, 405–408.
Cummings, J. (2017). Disease modification and Neuroprotection in neurodegenerative disorders. Transl. Neurodegener. *6*, 25.
Deffur, A., Wilkinson, R.J., Mayosi, B.M., and Mulder, N.M. (2018). ANIMA: Association network integration for multiscale analysis. Wellcome Open Res. *3*, 27.
Do-Ha, D., Buskila, Y., and Ooi, L. (2018). Impairments in Motor Neurons, Interneurons and Astrocytes Contribute to Hyperexcitability in ALS: Underlying Mechanisms and Paths to Therapy. Mol. Neurobiol. *55*, 1410–1418.
Dugger, B.N., and Dickson, D.W. (2017). Pathology of Neurodegenerative Diseases. Cold Spring Harb. Perspect. Biol. *9*, a028035.
Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., et al. (2013). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics *9*, 44–66.
Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.
Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46*, D649-55.
Farrar, M.A., Park, S.B., Vucic, S., Carey, K.A., Turner, B.J., Gillingwater, T.H., Swoboda, K.J., and Kiernan, M.C. (2017). Emerging therapies and challenges in spinal muscular atrophy. Ann. Neurol. *81*, 355–368.
Gaulton, A., Hersey, A., Nowotka, M.L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrian-Uhalte, E., et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res. *45*, D945-54.
Giuliani, A. (2017). The application of principal component analysis to drug discovery and biomedical data. Drug Discov. Today *22*, 1069–1076.
Gundry, R.L., White, M.Y., Murray, C.I., Kane, L.A., Fu, Q., Stanley, B.A., and Van Eyk, J.E. (2009). Preparation of Proteins and Peptides for Mass Spectrometry Analysis in a Bottom-Up Proteomics Workflow. Curr. Protoc. Mol. Biol. *88*, 10.25.1-10.25.23.
Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E., Logroscino, G., Robberecht, W., Shaw, P., Simmons, Z., and van den Berg, L. (2017). Amyotrophic lateral sclerosis. Nat. Rev. Dis. Prim. *3*, 17071.
Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive

chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.
Jenuwein, T. (2001). Translating the Histone Code. Science (80-.). *293*, 1074–1080.
Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R., and Keun, H.C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics *27*, 2917–2918.
Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353-61.
Kedaigle, A., and Fraenkel, E. (2018). Turning omics data into therapeutic insights. Curr. Opin. Pharmacol. *42*, 95–101.
Keenan, A.B., Jenkins, S.L., Jagodnik, K.M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A.B., Silverstein, M.C., Lachmann, A., et al. (2018). The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst. *6*, 13–24.
Kemmeren, P., Sameith, K., van de Pasch, L.A.L., Benschop, J.J., Lenstra, T.L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C.W., et al. (2014). Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors. Cell *157*, 740–752.
Kim, H.J., and Taylor, J.P. (2017). Lost in Transportation: Nucleocytoplasmic Transport Defects in ALS and Other Neurodegenerative Diseases. Neuron *96*, 285–297.
Kiriyama, Y., and Nochi, H. (2015). The Function of Autophagy in Neurodegenerative Diseases. Int. J. Mol. Sci. *16*, 26797–26812.
Kumar, A., Kumar Singh, S., Kumar, V., Kumar, D., Agarwal, S., and Rana, M.K. (2015). Huntington's disease: An update of therapeutic strategies. Gene *556*, 91–97.
Kumar, V., Sami, N., Kashav, T., Islam, A., Ahmad, F., and Hassan, M.I. (2016). Protein aggregation and neurodegenerative diseases: From theory to therapy. Eur. J. Med. Chem. *124*, 1105–1120.
Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
Labbadia, J., and Morimoto, R.I. (2013). Huntington's disease: Underlying molecular mechanisms and emerging concepts. Trends Biochem. Sci. *38*, 378–385.
Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science (80-.). *313*, 1929–1935.
Leiserson, M.D.M., Vandin, F., Wu, H.-T., Dobson, J.R., Eldridge, J. V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. *47*, 106–114.
Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A.T.M., Jørgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K., et al. (2007). Systematic Discovery of In Vivo Phosphorylation Networks. Cell *129*, 1415–1426.
Litichevskiy, L., Peckner, R., Abelin, J.G., Asiedu, J.K., Creech, A.L., Davis, J.F., Davison, D., Dunning, C.M., Egertson, J.D., Egri, S., et al. (2018). A Library of

Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations. Cell Syst. *6*, 424–43.e7.
Liu, X., and Locasale, J.W. (2017). Metabolomics: A Primer. Trends Biochem. Sci. *42*, 274–284.
Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. PLOS Comput. Biol. *13*, e1005457.
MacLachlan, G., Bean, R., and Ng, S. (2017). Clustering. Methods Mol. Biol. *1526*, 345–362.
Maharshi, V., and Hasan, S. (2017). Nusinersen: The First Option Beyond Supportive Care for Spinal Muscular Atrophy. Clin. Drug Investig. *37*, 807–817.
McAlister, G.C., Nusinow, D.P., Jedrychowski, M.P., Wühr, M., Huttlin, E.L., Erickson, B.K., Rad, R., Haas, W., and Gygi, S.P. (2014). MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. Anal. Chem. *86*, 7150–7158.
McColgan, P., and Tabrizi, S.J. (2018). Huntington's disease: a clinical review. Eur. J. Neurol. *25*, 24–34.
Medina, M.Á. (2013). Systems biology for molecular life sciences and its impact in biomedicine. Cell. Mol. Life Sci. *70*, 1035–1053.
Mohs, R.C., and Greig, N.H. (2017). Drug discovery and development: Role of basic biological research. Alzheimer's Dement. Transl. Res. Clin. Interv. *3*, 651–657.
Mulas, F., Li, A., Sherr, D.H., and Monti, S. (2017). Network-based analysis of transcriptional profiles from chemical perturbations experiments. BMC Bioinformatics *18*, 130.
O'Geen, H., Echipare, L., and Farnham, P.J. (2011). Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. Methods Mol. Biol. *791*, 265–286.
Oliveira, F.H.M., Machado, A.R.P., and Andrade, A.O. (2018). On the Use of t - Distributed Stochastic Neighbor Embedding for Data Visualization and Classification of Individuals with Parkinson's Disease. Comput. Math. Methods Med. *2018*, 1–17.
Parikshak, N.N., Gandal, M.J., and Geschwind, D.H. (2015). Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nat. Rev. Genet. *16*, 441–458.
Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.
Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C.B., White, F.M., Saghatelian, A., and Fraenkel, E. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. Nat. Methods *13*, 770–776.
Pozhitkov, A.E., Tautz, D., and Noble, P.A. (2007). Oligonucleotide microarrays: widely applied poorly understood. Briefings Funct. Genomics Proteomics *6*, 141–148.
Prathipati, P., and Mizuguchi, K. (2015). Systems Biology Approaches to a Rational Drug Discovery Paradigm. Curr. Top. Med. Chem. *16*, 1009–1025.
Rees, M.G., Seashore-Ludlow, B., Cheah, J.H., Adams, D.J., Price, E. V., Gill, S., Javaid, S., Coletti, M.E., Jones, V.L., Bodycombe, N.E., et al. (2016). Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat. Chem. Biol. *12*, 109–116.
Reyna, M.A., Leiserson, M.D.M., and Raphael, B.J. (2018). Hierarchical HotNet:

identifying hierarchies of altered subnetworks. Bioinformatics *34*, i972-80.
Ritz, A., Poirel, C.L., Tegge, A.N., Sharp, N., Simmons, K., Powell, A., Kale, S.D., and Murali, T. (2016). Pathways on demand: automated reconstruction of human signaling networks. Npj Syst. Biol. Appl. *2*, 16002.
Ross, C. a (2002). Polyglutamine Pathogenesis : Minireview Emergence of Unifying Mechanisms for Huntington ' s Disease and Related Disorders. *35*, 819–822.
Ross, C. a., and Tabrizi, S.J. (2011). Huntington's disease: From molecular pathogenesis to clinical treatment. Lancet Neurol. *10*, 83–98.
Schneider, M.V. (2013). Defining systems biology: A brief overview of the term and field. Defining Systems Biology: A Brief Overview of the Term and Field. In: Schneider M. (eds) In Silico Systems Biology. Methods Mol. Biol. (Methods Protoc. *1021*, 1–11.
Schulte, J., and Littleton, J.T. (2011). The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. Curr. Trends Neurol. *5*, 65–78.
Si, Y., Liu, P., Li, P., and Brutnell, T.P. (2014). Model-based clustering for RNA-seq data. Bioinformatics *30*, 197–205.
Simeoni, C., Dinicola, S., Cucina, A., Mascia, C., and Bizzarri, M. (2018). Systems Biology Approach and Mathematical Modeling for Analyzing Phase-Space Switch During Epithelial-Mesenchymal Transition. Methods Mol. Biol. *1702*, 95–123.
Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. *46*, D661-7.
Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437–52.e17.
Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.
Tulloch, L.B., Menzies, S.K., Coron, R.P., Roberts, M.D., Florence, G.J., and Smith, T.K. (2018). Direct and indirect approaches to identify drug modes of action. IUBMB Life *70*, 9–22.
Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S.C., Chayes, J., Borgs, C., Zecchina, R., and Fraenkel, E. (2013). Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner Forest Problem. J. Comput. Biol. *20*, 124–136.
Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput. Biol. *12*, e1004879.
Varma, H., Lo, D., and Stockwell, B. (2008). High Throughput Screening for Neurodegeneration and Complex Disease Phenotypes. Comb. Chem. High Throughput Screen. *11*, 238–248.
Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237-45.
Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for

transcriptomics. Nat. Rev. Genet. *10*, 57–63.
Webber, J.T., Kaushik, S., and Bandyopadhyay, S. (2018). Integration of Tumor Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics. Cell Syst. *7*, 526–36.e6.
Wehling, M. (2009). Assessing the translatability of drug projects: What needs to be scored to predict success? Nat. Rev. Drug Discov. *8*, 541–546.
Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074-82.
Wong, C.H., Siah, K.W., and Lo, A.W. (2018). Estimation of clinical trial success rates and related parameters. Biostatistics *20*, 273–286.
Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Rodríguez Martínez, M., López, G., Mattioli, M., Realubit, R., et al. (2015). Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell *162*, 441–451.
Wood, L.B., Winslow, A.R., and Strasser, S.D. (2015). Systems biology of neurodegenerative diseases. Integr. Biol. *7*, 758–775.
Wynn, M.L., Consul, N., Merajver, S.D., and Schnell, S. (2012). Logic-based models in systems biology: a predictive and parameter-free network analysis method. Integr. Biol. *4*, 1323–1337.
Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955-61.
Zampieri, M., Szappanos, B., Buchieri, M.V., Trauner, A., Piazza, I., Picotti, P., Gagneux, S., Borrell, S., Gicquel, B., Lelievre, J., et al. (2018). High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. Sci. Transl. Med. *10*.
Zuccato, C., Valenza, M., and Cattaneo, E. (2010). Molecular Mechanisms and Potential Therapeutical Targets in Huntington ' s Disease. Physiol Rev *90*, 905–981.

Chapter 2: An Interpretable Machine Learning Model Reveals Disease-Relevant Modes of Action of Small Molecules

This work has been submitted for publication.

Natasha L Patel-Murray, Miriam Adam, Nhan Huynh, Brook T. Wassie, Pamela Milani, and Ernest Fraenkel.

As part of this work, I would like to gratefully acknowledge members of the MIT BioMicro Center, the Whitehead Institute Metabolite Profiling Core Facility, the Thermo Fisher Scientific Center for Multiplexed Proteomics at Harvard Medical School, the Koch Institute Swanson Biotechnology Center High Throughput Screening Facility, and Dr. David Sabatini's lab at the Whitehead Institute for assistance with sequencing data collection, metabolomic data collection, proteomic data collection, imaging, and bioenergetics data collection, respectively.

My contributions:
I performed the experiments, some with assistance, and the analysis of the different data. I created the figures and wrote the manuscript.

2.1 Abstract

High-throughput screening and gene signature analyses frequently identify lead therapeutic compounds with unknown modes of action (MoAs), and the resulting uncertainties can lead to the failure of clinical trials. We developed a multi-omics approach for uncovering MoAs through an interpretable machine learning model of the transcriptomic, epigenomic, metabolomic, and proteomic effects of compounds (Figure 2-1). We applied this approach to examine compounds with beneficial effects in models of Huntington's disease, finding common MoAs for previously unrelated compounds that were not predicted based on similarities in the compounds' structures, connectivity scores, or binding targets. We experimentally validated two such disease-relevant MoAs, autophagy activation and bioenergetics manipulation. Our interpretable machine learning approach can be used to find and evaluate MoAs in future drug development efforts.

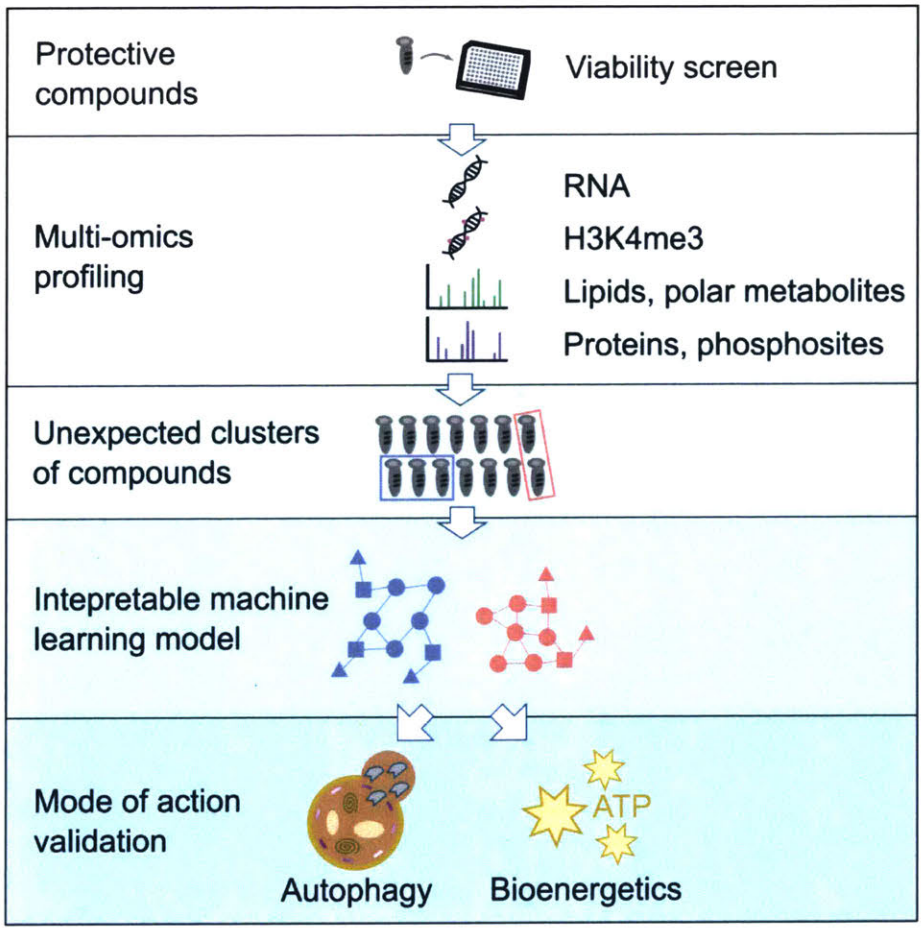

Figure 2-1 Graphical Abstract

2.2 Introduction

Unknown modes of action of drug candidates can lead to unpredicted consequences on effectiveness and safety. Computational methods, such as the analysis of gene signatures, and high-throughput experimental methods have accelerated the discovery of lead compounds that affect a specific target or phenotype (Lamb et al., 2006; Litichevskiy et al., 2018; Subramanian et al., 2017). However, these advances have had not dramatically changed the rate of drug approvals. Between 2000 and 2015, 86% of drug candidates failed to earn FDA-approval, with toxicity or a lack of efficacy being common reasons for their clinical trial termination (Wehling, 2009; Wong et al., 2018). Even compounds identified for binding to a specific target can have complex downstream functional consequences, or modes of action (MoAs) (Tulloch et al., 2018). Understanding the MoAs of compounds remains a crucial challenge in increasing the success rate of clinical trials and drug-repurposing efforts (Tulloch et al., 2018; Wehling, 2009).

Computational approaches have contributed to the discovery of MoAs. Using the Connectivity Map data, tools like MANTRA can predict MoAs of new compounds based on their gene expression similarity to reference compounds with known MoAs (Iorio et al., 2013). To combat antibiotic resistance, reference compounds were also used to infer MoAs of uncharacterized antimicrobial compounds by comparing their untargeted metabolomic profiles in bacteria (Zampieri et al., 2018). From human cancer cell lines, basal gene expression signatures were correlated with sensitivity patterns of compounds to identify previously unknown activation mechanisms and compound binding targets (Rees et al., 2016). Similarly, gene expression profiles of human lymphoma cells treated with anti-cancer drugs were compared using the gene regulatory network-based DeMAND algorithm to predict novel targets and unexpected similarities between the drugs (Woo et al., 2015). However, all of these methods require prior context-specific knowledge, such as data from reference compounds with known MoAs, sensitivity data, or gene-regulatory interactions.

More general approaches to discover MoAs are urgently needed. In the context of late-onset neurodegenerative disorders like Huntington's Disease (HD), screening efforts focused on protein aggregation, neuronal death, and caspase activation

phenotypes have found many compounds that have disease-altering potential, but none have been successful in clinical trials (Varma et al., 2008). HD is an autosomal dominant, fatal neurodegenerative disorder that results in massive striatal neuronal cell death (Kumar et al., 2015). The disease is caused by a trinucleotide repeat expansion in the huntingtin gene, which encodes an expanded polyglutamine domain in the huntingtin protein (Kumar et al., 2015). Although the exact function of huntingtin is unclear, it has been shown to interact with many proteins and to be involved in transcription, anti-apoptotic activity, and the trafficking processes of vesicles and organelles (Schulte and Littleton, 2011). Within brain cells, mutant huntingtin causes transcriptional dysregulation, impaired cytoskeletal motor functions, compromised energy metabolism, and abnormal immune activation (Schulte and Littleton, 2011).

Over the years, many compounds have been discovered that confer a protective effect in HD model systems (Zuccato et al., 2010). In some cases, direct binding targets are known, but these may not always be in the therapeutic pathway. A study using a small molecule sphingolipid enzyme inhibitor, for example, found a novel MoA related to histone acetylation through the analysis of gene expression and epigenetic profiles in the murine STHdh^{Q111} HD cell model (Pirhaji et al., 2017). As all small-molecule therapeutics have so far failed to modify HD in clinical trials, understanding the disease-relevant MoAs is critical to guide future therapeutic approaches that could target these pathways with new molecules.

We reasoned that the discovery of MoAs must begin with an unbiased approach. Some compounds may have largely transcriptional effects, while others may primarily impact signaling or metabolism. With improvements in omics technology, it is now possible to systematically assess each of these areas. Technologies such as RNA-Seq, ChIP-Seq, and mass spectrometry provide extensive measurements of gene expression, chromatin accessibility, metabolite expression, protein expression, and post-translational modifications. The integration of these omics data can provide a more comprehensive view of the compounds and allow for discoveries that could be overlooked in the analysis of any individual dataset (Kedaigle and Fraenkel, 2018).

To systematically reveal disease-relevant MoAs, we developed a multi-omics machine learning approach (Figure 2-2) that does not require context-specific prior

knowledge or reference compounds. We used a hierarchical data generation strategy and began with a set of compounds previously reported to alleviate an HD phenotype in at least one HD model system. We filtered the compounds using a viability assay to find those that are protective in the well-established murine striatal STHdh^{Q111} HD cell model. We then profiled compound-treated cells using transcriptomics and untargeted metabolomics. Interestingly, we show that previously unrelated compounds cluster together based on their molecular profiles. For two interesting clusters of compounds, we then gathered proteomic data and epigenomic data.

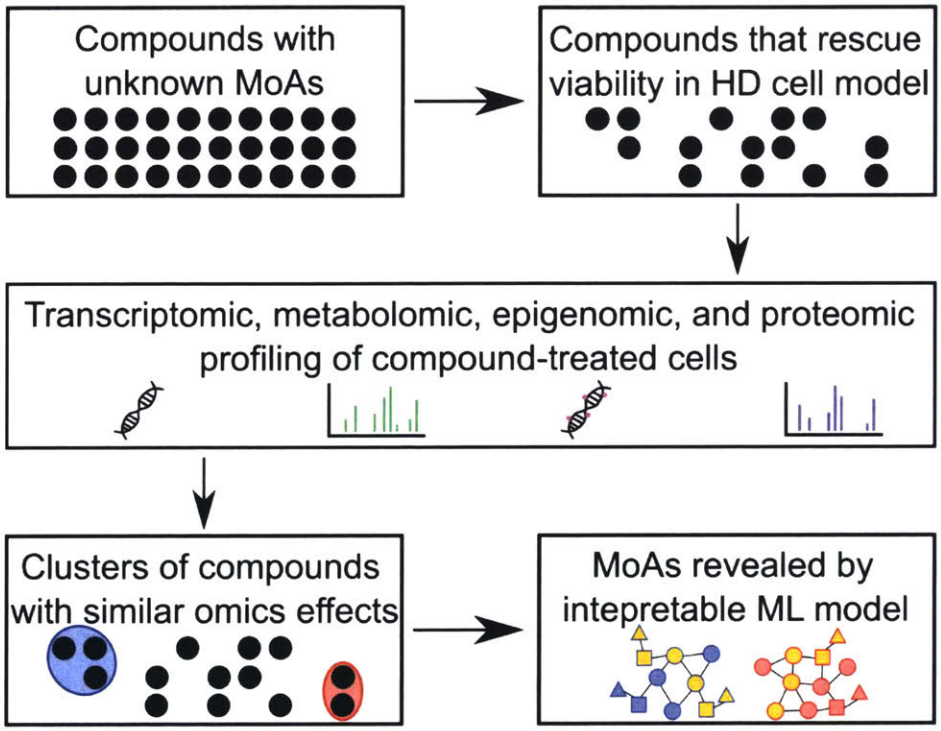

Figure 2-2 General workflow of study.

Compounds with unknown MoAs were found to be protective in HD cells. After multi-omics profiling, groups of protective compounds were shown to cluster together. An interpretable machine learning (ML) model revealed compounds' MoAs, which were validated experimentally.

To reveal the MoAs for these compounds we applied an interpretable machine-learning algorithm. We mapped each type of molecular data to a network of molecular interactions. Network optimization of this large interactome highlights the functional changes induced by the compounds. This approach prioritized two disease-relevant processes, autophagy activation and mitochondrial respiration inhibition, as key MoAs of a subset of these compounds. Through cellular imaging, biochemical, and energetics assays, we confirmed these MoAs in the STHdh^{Q111} murine model. We also demonstrated that these effects on autophagy are reproducible across species and across cell types. To our knowledge, this is the first report of using machine learning on transcriptomic, epigenomic, metabolomic, and proteomic data to profile compounds' effects and determine disease-relevant MoAs. Our results and multi-omics network approach can be used to guide targeted drug selection and repurposing in HD and other disease contexts.

2.3 Results

2.3.1 Cell Viability Assay Categorizes Compounds by Protectiveness

More than 100 compounds were previously reported to reverse a disease phenotype in at least one HD model system (Bates et al., 2014). We examined 30 of these compounds that were commercially available (Table 2-S1), and determined their protectiveness in the well-established STHdh cell culture model of HD. These murine striatal neuronal progenitor cells express the polyglutamine-expanded (STHdh^{Q111}) or wild-type (STHdh^{Q7}) human huntingtin gene (Trettel et al., 2000). As has been previously reported, STHdh^{Q111} and STHdh^{Q7} cells differ in their sensitivity to serum deprivation (Trettel et al., 2000). As a result, we tested the ability of compounds to extend the viability of STHdh^{Q111} cells under these conditions. Of the compounds, 14 were significantly protective (p-value < 0.001) when compared to the STHdh^{Q111} vehicle control (Figure 2-3A). The remaining 16 compounds either did not significantly decrease cell death or were toxic to the cells at all tested concentrations.

2.3.2 Molecular Profiles Reveal Unexpected Similarities between Compounds

To assess the compounds' molecular effects on transcription and metabolism, we performed RNA-Seq and untargeted metabolite profiling on STHdh^{Q111} cells treated with the 14 protective compounds and vehicle control, in triplicate. We also included the STHdh^{Q7} vehicle control for comparison. We measured the levels of 18,178 genes, 1,530 untargeted lipids, and 1,805 untargeted polar metabolites in all samples. In most of the compound-treated samples, we found thousands of statistically significant differentially expressed genes (FDR-adjusted p-value < 0.05) compared to the STHdh^{Q111} vehicle control (Figure 2-3B). Though some compounds affected several hundred measured metabolites, many of the compounds had little effect on the lipids and polar metabolites (Figure 2-3B).

To reveal similarities between the compounds' profiles, we clustered the RNA, lipid, and polar metabolite data separately (Figure 2-4A). In the gene expression data, five compounds reproducibly clustered tightly together in a group distinct from the STHdh^{Q111} vehicle control samples. Although these compounds formed only one distinct group in the gene expression data, they separated into two distinct groups in the metabolite profiling data. Cyproheptadine, loxapine, and pizotifen form Group A and were previously shown to block caspase activation and increase ERK activation (Sarantos et al., 2012). Group B, surprisingly, consists of the previously unrelated compounds diacylglycerol kinase inhibitor II (DKI) and meclizine. Some compounds, such as 4-deoxypyrididoxine (DOP) and cysteamine, can be separated from the STHdh^{Q111} vehicle control samples only in the metabolite data, but do not cluster tightly with other compounds. Compounds that clustered together did not have the most similar structures, calculated using the maximum common substructure Tanimoto coefficients in ChemMine tools (Figure 2-S2) (Backman et al., 2011). Likewise, compound pairs with the strongest connectivity scores, as reported by the Connectivity Map using their L1000 gene expression data, did not cluster together in the omics data (Figure 2-S3) (Subramanian et al., 2017).

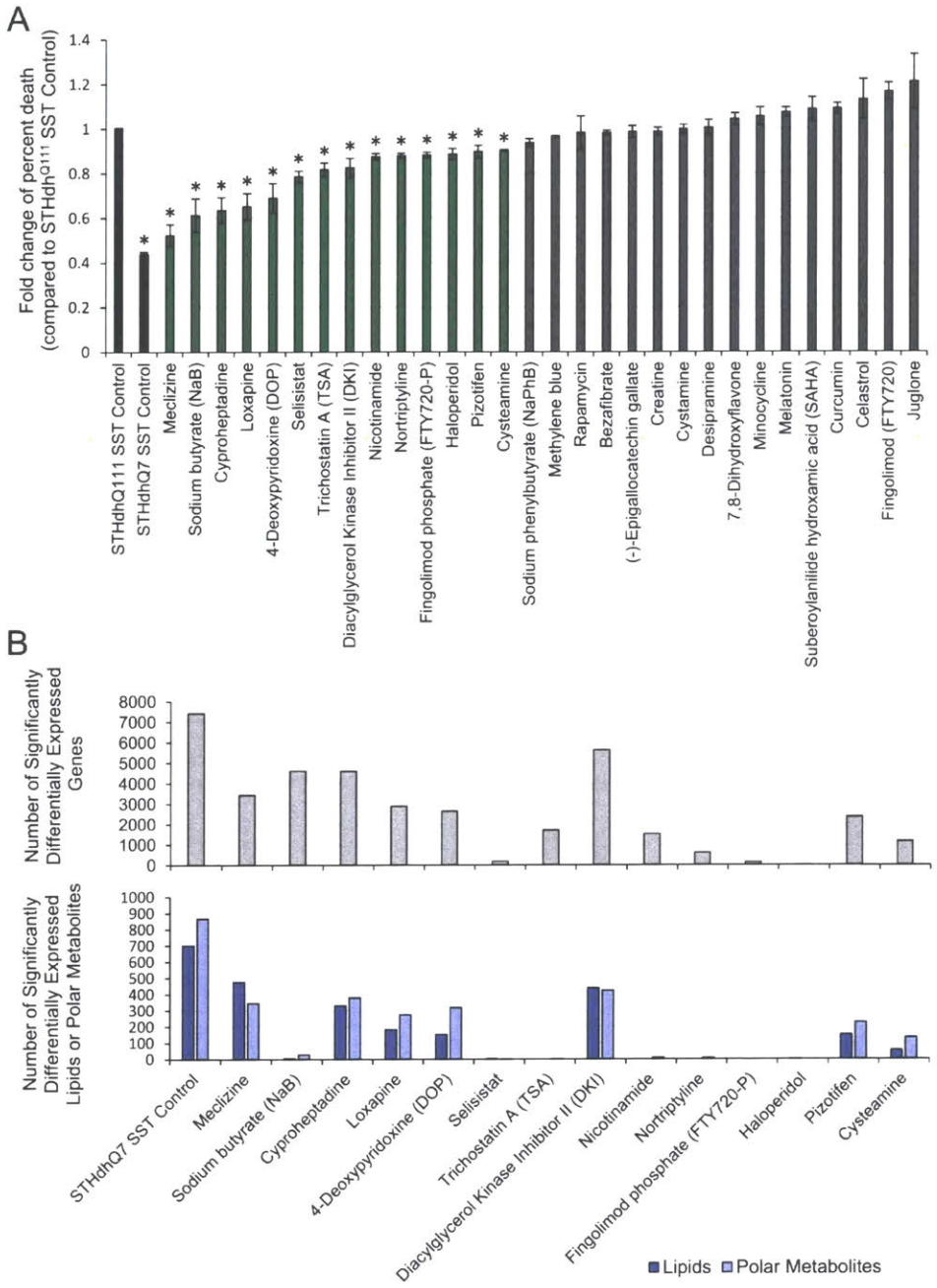

Figure 2-3 Compounds have diverse effects on viability, gene expression, and metabolite expression in the STHdh^{Q111} cell model.

(A) Cell viability assay categorizes 14 compounds as protective and 16 as unprotective. Data are represented as mean ± SD. *p-value < 0.001. The black, green, and gray bars indicate controls, protective compounds, and unprotective compounds, respectively.

(B) The number of transcriptomic and metabolic changes in compound-treated cells compared to controls.

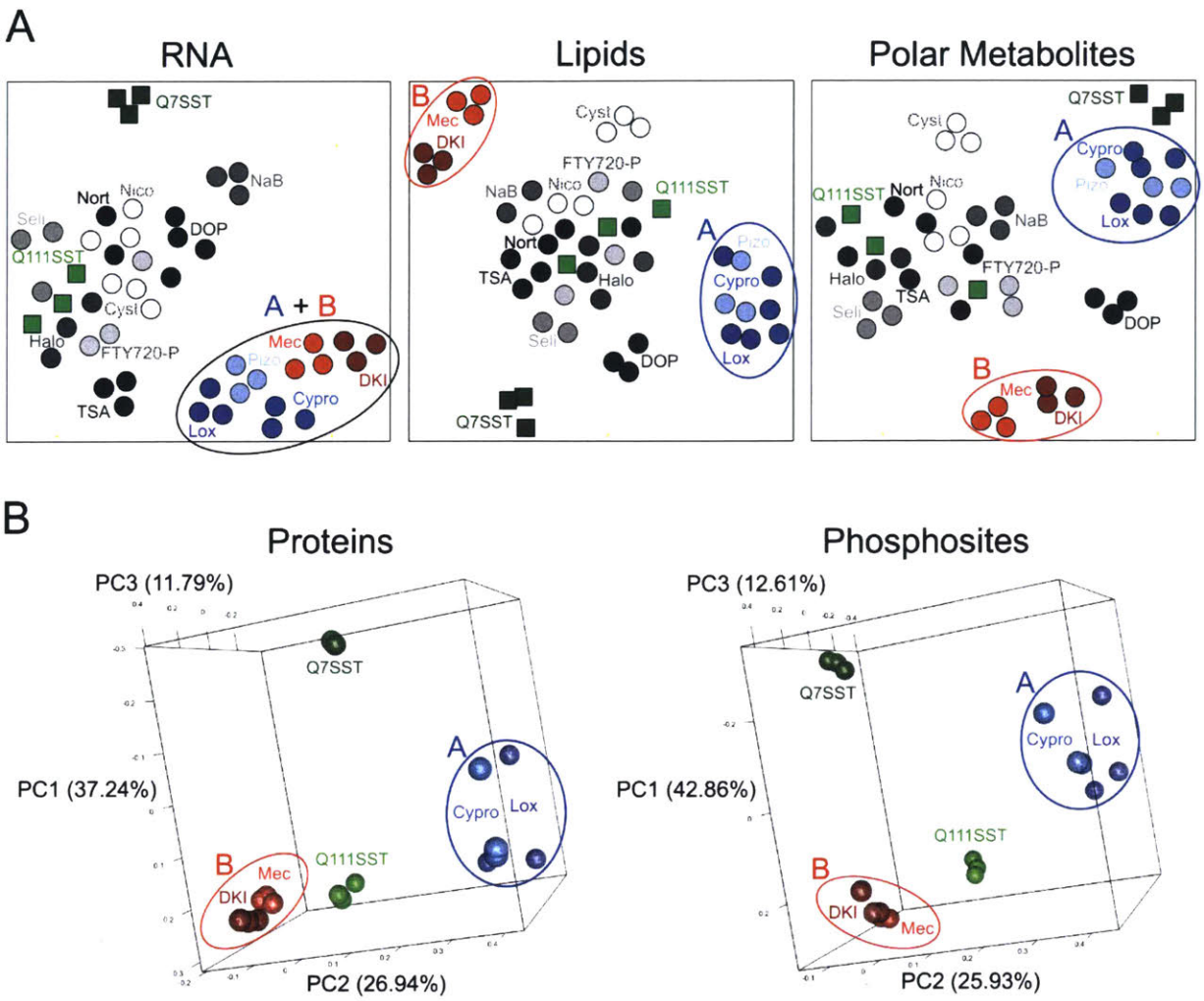

Figure 2-4 Omics profiles reveal unexpected similarities between compounds.

(A) Clustering of metabolite profiling data reveals two distinct groups of compounds that are inseparable in the gene expression data, as displayed in t-SNE plots. The blue and red ellipses indicate the Group A and Group B compounds, respectively. Q7SST = STHdh^{Q7} SST control; Q111SST = STHdh^{Q111} SST control; Mec = meclizine; NaB = sodium butyrate; Cypro = cyproheptadine; Lox = loxapine; DOP = 4-Deoxypyridoxine; Seli = selisistat; TSA = trichostatin A; DKI = diacylglycerol kinase inhibitor II; Nico = nicotinamide; Nort = nortriptyline; FTY720-P = fingolimod phosphate; Halo = haloperidol; Pizo = pizotifen; Cyst = cysteamine.

(B) Clustering of proteomics data, as shown in three-dimensional PCA plots.

See also Figures 2-S1-S3, Tables 2-S2-S6.

To further characterize the compounds in Groups A and B, we performed global proteomics and phosphoproteomics analysis. We identified and measured the levels of 6,281 proteins and 2,560 phosphosites in controls and compound-treated cells. We selected two compounds from Group A, cyproheptadine and loxapine, and the two compounds in Group B because they had the most RNA and metabolite changes compared to the STHdh^{Q111} vehicle controls. These four compounds show several statistically significantly differentially expressed proteins and phosphosites, and they also cluster reproducibly by their respective groups in both types of proteomic data (Figure 2-4B). The differential genes and proteins of the Group A compounds are significantly enriched (FDR-adjusted p-value < 0.05) in 882 and 2 gene ontology (GO) processes, respectively (Tables 2-S2, 2-S4). The Group B differential genes are significantly enriched (FDR-adjusted p-value < 0.05) in 911 GO processes, but the Group B differential proteins have no significant GO process enrichment (Table 2-S5). Using the IMPaLA tool for metabolite pathway analysis, the Group A and Group B differential metabolites are significantly enriched (FDR-adjusted p-value < 0.05) in 82 and 42 pathways, respectively (Tables 2-S3, 2-S6).

2.3.3 Machine Learning Network Models Prioritize HD-Relevant Modes of Action

Analyzed separately, the omics data provide a confusing perspective of the changes associated with each compound, pointing to hundreds of potential pathways and processes. To develop a comprehensive view of the compounds' downstream effects, we turned to dimensionality reduction approaches that leverage known molecular interactions. PIUMet and Omics Integrator use network optimization to identify a subset of the input features that can be linked to each other through direct or indirect molecular interactions (Pirhaji et al., 2016; Tuncbag et al., 2016). We first applied PIUMet to map untargeted metabolomics to the interactome.

To identify the regulatory factors driving changes in transcription, we profiled the H3K4me3 epigenetic modification, which is associated with promoter regions in accessible chromatin, using ChIP-Seq (Heintzman et al., 2007). Though we found few differential peaks between STHdh^{Q111} control cells and compound-treated cells, we used the overall epigenetic signature as a measurement of transcription factor binding

accessibility. We predicted transcription factors using a motif analysis approach applied to the differentially expressed genes and the H3K4me3 regions.

We then applied Omics Integrator for graph-constrained dimensionality reduction. The inputs were the differential metabolites, proteins, phosphoproteins, and predicted transcriptional regulators for each of the two compound groups. After filtering the networks based on node robustness and specificity, we found significant GO enrichment for pathways relevant in HD. The Group A network was highly enriched for the autophagy, protein localization and transport, and cytoskeleton organization processes (Table 2-S7). The Group B network was highly enriched for the mitochondrial electron transport, sterol metabolism, and amino acid processes (Table 2-S8). Based on the network enrichment, we prioritized the autophagy and mitochondrial respiration pathways for further experimental testing (Figure 2-5A-B).

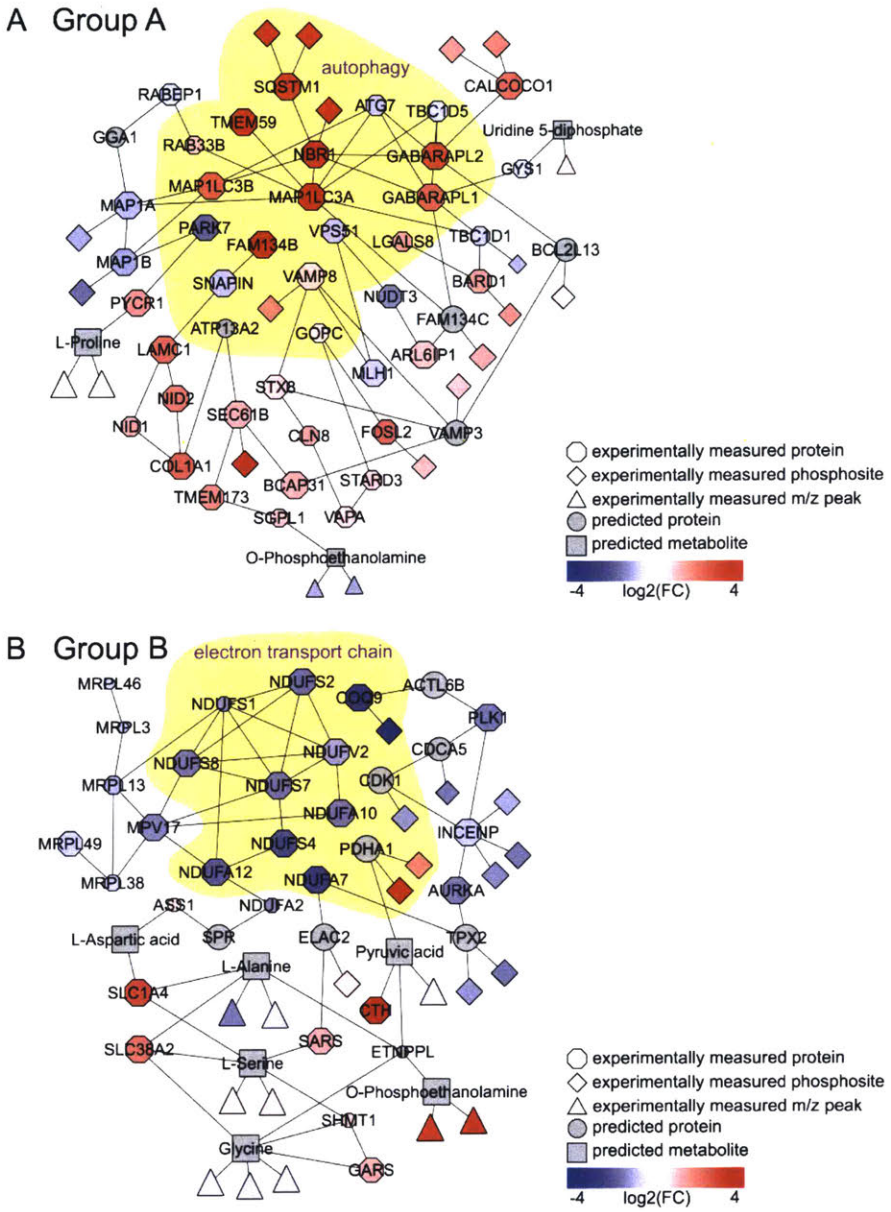

Figure 2-5 Machine learning network models prioritize HD-relevant pathways.

(A) The autophagy pathway is significantly enriched (p-value < 0.05) in the Group A compound network, a subnetwork of which is shown. The highlighted yellow region indicates those proteins that are part of the autophagy GO term.

(B) The electron transport chain is significantly enriched (p-value < 0.05) in the Group B compound network, a subnetwork of which is shown. The highlighted yellow region indicates those proteins that are part of the electron transport chain GO term, part of the mitochondrial respiration pathway.

See also Tables 2-S7, 2-S8.

2.3.4 Autophagy is Up-Regulated by Group A Compounds

Autophagy, which appeared in the optimized networks, is also known to be dysregulated in HD (Martin et al., 2015). We measured the levels of autophagic vacuoles in the STHdh^{Q111} cells using fluorescent staining. We found that the Group A compounds significantly increased the fluorescence intensity, indicating an increase in the number of autophagic vacuoles (Figure 2-6A). To further quantify autophagy differences, we examined levels of microtubule-associated protein light chain 3 (LC3), which is widely used to monitor autophagy (Mizushima and Yoshimori, 2007). We quantified the levels of LC3-II and LC3-I by Western blots in control and compound-treated cells. We found a significant increase in the LC3-II to LC3-I ratio with treatment of the Group A compounds, but no significant change with treatment of the Group B compounds (Figure 2-6B-C), indicating that the Group A compounds increase formation of autophagic vacuoles. To determine whether this increase was due to an activation of autophagy or a degradation blockage of the autophagic vacuoles, we treated the cells with and without bafilomycin A1 (bafA), an inhibitor of late-stage autophagy (Mizushima and Yoshimori, 2007). We found a further increase in the LC3-II to LC3-I ratio in all of the conditions upon treatment of bafA, indicating that the Group A compounds activate autophagy in the STHdh^{Q111} cells.

As STHdh^{Q111} cells derive from a mouse model of HD, we also tested whether the MoA was relevant in human cells. In human, neuronal SH-SY5Y cells, the fluorescent staining assay showed an increase in the number of autophagic vacuoles in the Group A compound-treated cells compared to control cells (Figure 2-S4). Similar results were obtained in HEK293 cells, which are also human but non-neuronal. In both cell types, the Group A compounds significantly increase the LC3-II to LC3-I ratio, while the Group B compounds do not significantly change the ratio (Figure 2-7A-D). The addition of bafA further increased the ratio in all conditions, indicating an activation of autophagy by the Group A compounds in all three cell types.

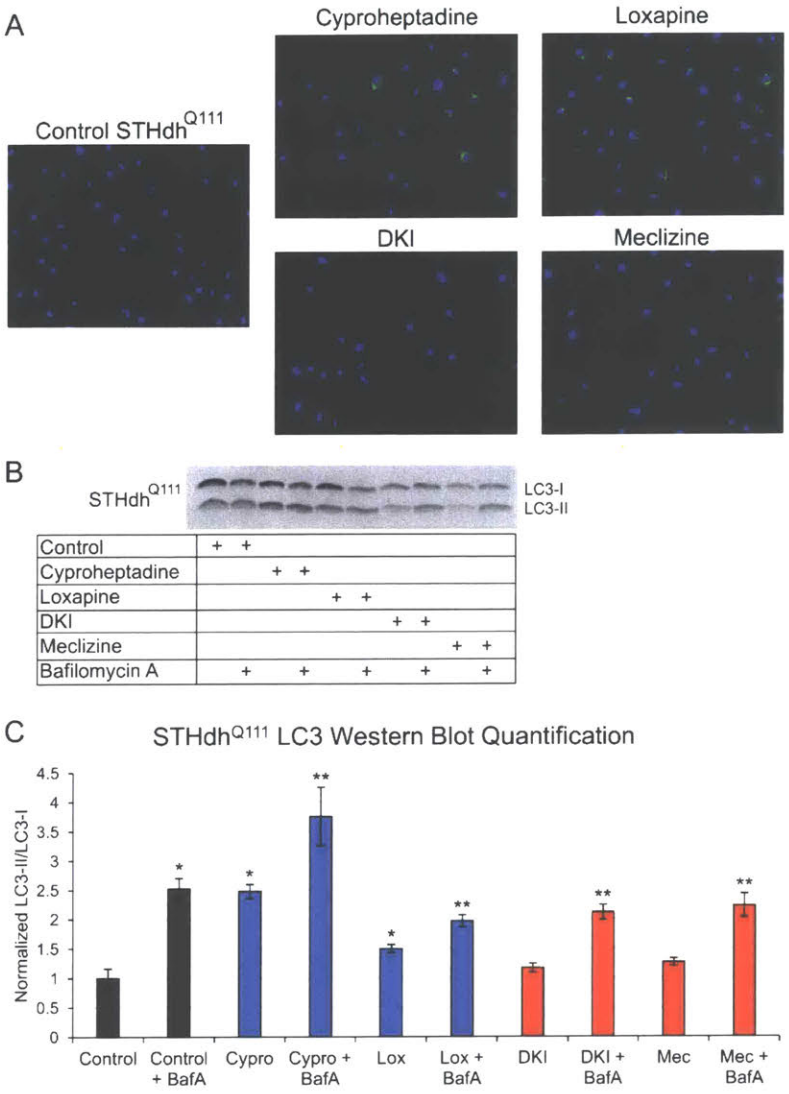

Figure 2-6 Autophagy is up-regulated by Group A compounds in murine STHdh^{Q111} cells.

(A) Fluorescent staining of autophagic vacuoles in Group A compound-treated cells compared to Group B compound-treated or control cells. Blue fluorescence indicates nuclei and green fluorescence indicates autophagic vacuoles.

(B) A representative western blot showing LC3-II and LC3-I levels to determine how the compounds affect autophagy. BafA was used to determine whether the compounds activate autophagy or inhibit vacuole degradation.

(C) Quantification of the LC3-II to LC3-I ratio normalized to the control from the western blot. Data are represented as mean ± SD. *p-value < 0.05 compared to Control; **p-value < 0.05 compared to condition-matched non-bafA treatment.

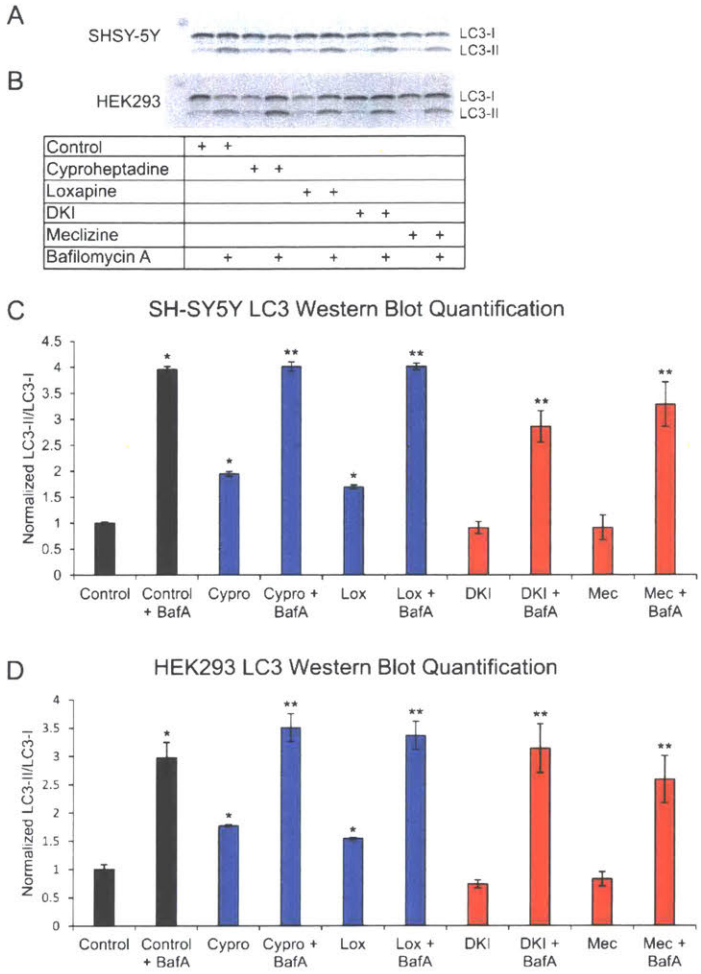

Figure 2-7 Autophagy is up-regulated by Group A compounds in human SH-SY5Y and HEK293 cells.

(A) A representative western blot showing LC3-II and LC3-I levels to determine how the compounds affect autophagy in SH-SY5Y cells.

(B) A representative western blot showing LC3-II and LC3-I levels to determine how the compounds affect autophagy in HEK293 cells.

(C) Quantification of the LC3-II to LC3-I ratio normalized to the control in SH-SY5Y cells from the western blot. *p-value < 0.05 compared to Control; **p-value < 0.05 compared to condition-matched non-bafA treatment.

(D) Quantification of LC3-II to LC3-I ratio normalized to the control in HEK293 cells from the western blot. Data are represented as mean ± SD. *p-value < 0.05 compared to Control; **p-value < 0.05 compared to condition-matched non-bafA treatment.

See also Figure 2-S4.

2.3.5 Bioenergetics are Altered Differently by Each Group of Compounds

The network analysis of Group B compounds suggested an MoA relating to bioenergetics, which are also known to be affected in HD (Kedaigle et al., 2019). To test this, we used the Seahorse Real-Time ATP Production assay to measure the rates of mitochondrial respiration and glycolysis in STHdh^{Q111} control and compound-treated cells. We found that both Group B compounds indeed inhibited mitochondrial respiration and enhanced glycolysis compared to the control cells, but the total ATP production levels were unchanged (Figure 2-8A-C). Interestingly, we also found significantly enhanced mitochondrial respiration and slightly enhanced glycolysis ATP production rates by the Group A compounds. The net ATP production was increased by the Group A compounds compared to the STHdh^{Q111} control cells. The two groups of compounds show seemingly opposite effects, where the Group A compounds primarily rescue the mitochondrial respiration deficit and the Group B compounds rescue the glycolysis deficit present in the STHdh^{Q111} cells compared to the STHdh^{Q7} cells.

2.4 Discussion

The molecular effects of drug candidates are complex and can be difficult to interpret. Cataloguing efforts, such as those by the Connectivity Map, LINCS and Genomics of Drug Sensitivity in Cancer consortia, have made it possible to rapidly compare small molecules using expression or bioactivity data (Gaulton et al., 2017; Lamb et al., 2006; Litichevskiy et al., 2018; Rees et al., 2016; Subramanian et al., 2017; Wishart et al., 2018a; Yang et al., 2013). In cases where a compound of interest shows similarities to one with known MoAs, this process can lead to functional insights. However, these compendia themselves contain thousands of compounds that do not match up to any reference.

Our findings demonstrate the value of an approach that combines multi-omics with an interpretable machine learning method to determine previously unknown MoAs, even in the absence of a comparable reference. Using this approach, we identified and experimentally validated Huntington's Disease-relevant MoAs for two classes of compounds.

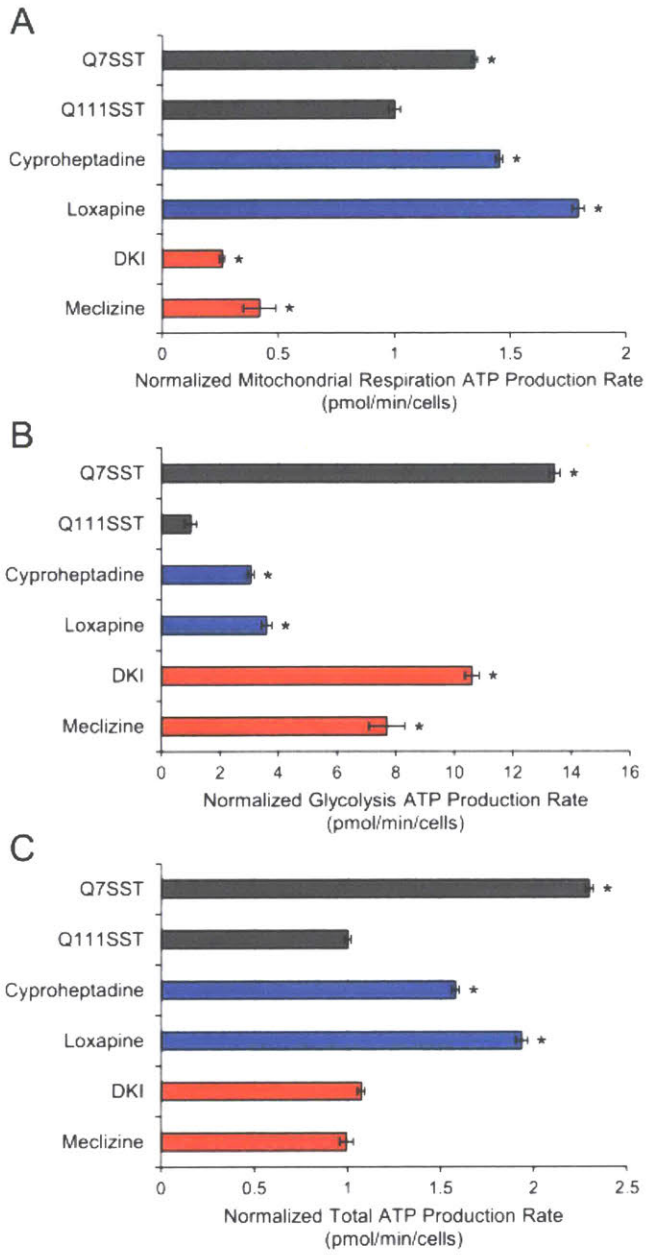

Figure 2-8 Bioenergetics are altered differently by Group A and Group B compounds in STHdh^{Q111} cells.

(A) Quantification of the mitochondrial respiration ATP production rate normalized to the STHdh^{Q111} control.

(B) Quantification of the glycolysis ATP production rate normalized to the STHdh^{Q111} control.

(C) Quantification of the total ATP production rate normalized to the STHdh^{Q111} control. Data are represented as mean ± SEM. *p-value < 0.05.

Although the 30 compounds tested in this study were previously shown to reverse an HD phenotype, their disease-relevant MoAs were unknown. Analyzing the 14 protective compounds, we found that unexpected compounds had similar molecular effects. These clusters of compounds would not have been predicted solely based on the compounds' phenotypic viability readouts, structural similarities, connectivity scores, or known binding targets (Figures 2-3A, 2-S2, 2-S3).

It is important to note that the disease-relevant MoA might be distinct from that previously reported in the literature. Cyproheptadine and meclizine are both antihistamines known to antagonize the histamine H1 receptor (Wishart et al., 2018a). Based on their reported MoAs and their effectiveness in HD models, these two antihistamines might have been expected to have similar therapeutic mechanisms. However, the two compounds have dissimilar omics profiles and fall in different clusters (Figure 2-4A). Indeed, our machine learning approach predicted that they would have different effects on autophagy and bioenergetics, which we confirmed experimentally. On the other hand, our approach suggested, and we experimentally confirmed, a common disease-relevant MoA for DKI and meclizine, whose reported targets (diacylglycerol kinase and histamine H1 receptor, respectively) are unrelated. Thus, the phenotypic effects of a compound can be unpredictable even when a direct target is known.

Taken together, the five different omics data types show the extensive changes that occur after compound treatment in the STHdh^{Q111} cells. Groups A and B affected a similar number of genes and are nearly indistinguishable in their RNA profiles (Figure 2-3A). For these compounds, metabolite profiling data proved more useful than the gene expression data in revealing their different effects. These distinct groups are reproduced in the proteomic data and ultimately reflect the functional differences in biological processes, such as in the autophagy and bioenergetics pathways. Metabolomic assays are less expensive than proteomic assays, but unlike the transcriptomic data, they still provide the resolution needed to suggest differences between the two groups of compounds. However, other compounds showed little to no effect on metabolites, but did robustly alter gene expression. It is also noteworthy that though the clusters of compounds were the same, the Group A compounds affected more proteins and

phosphosites than the Group B compounds, whereas the Group B compounds affected more lipids and polar metabolites than the Group A compounds (Figure 2-S1). Thus, there may be no single omics method that will provide sufficient data for all compounds.

Each omics data type highlights changes induced by the compounds, but an interpretable machine learning approach for dimensionality reduction prioritized cellular processes that connected the omics effects. The physical interaction networks allowed us to identify and prioritize the autophagy and mitochondrial respiration pathways as processes affected by the Group A and Group B compounds, respectively (Figure 2-5A-B). These pathways were not top hits in the gene, metabolite, or protein enrichments for either group of compounds (Tables 2-S2-S6). In each data type alone, there are hundreds to thousands of changes and no direct mechanistic insight, but physical interaction models enable identification of compounds' MoAs.

One limitation to our approach is that the networks have undirected edges and do not provide causal information for the highlighted processes. Instead, the pathways currently must be tested experimentally to ascertain the directionality of the changes. Using the network results to guide experimental efforts, we determined that autophagy is activated by Group A compounds and that mitochondrial respiration is inhibited by Group B compounds in the STHdh^{Q111} cells (Figures 2-6A-C, 2-8A-C). The specific effects on autophagy, mitochondrial respiration, and glycolysis by the Group A compounds were previously unknown. We verified that the Group A compounds activate autophagy in other cell lines, namely human SH-SY5Y and HEK293 cells (Figures 2-7A-D, 2-S2A-B). It has been reported that the Group B compound meclizine is an inhibitor of mitochondrial respiration and activator of glycolysis, which we confirmed (Gohil et al., 2013; Hong et al., 2016). The other Group B compound, DKI, has not previously been associated in changes in bioenergetics, but has the same effect as meclizine in the STHdh^{Q111} cells. Though our multi-omics machine learning approach can identify a compound's MoAs, it does not pinpoint the precise changes in the pathways required to produce the compound's effect. Future experimental efforts to modulate specific parts of the autophagy and bioenergetic pathways could lead to an increased understanding of the compounds' effects.

The omics data we collected in this study are available and can be used to guide drug repurposing efforts. For example, we hypothesize that the Group A compounds could be strong drug candidates for diseases where autophagy is deficient, like in neurodegenerative diseases. Similarly, the compounds we profiled can be used as reference compounds for existing methods of MoA identification. For instance, if the omics profiles for a new compound are comparable to those in Group B, then it suggests that the new compound would also affect bioenergetics. Overall, our multi-omics machine learning approach can be used to find and evaluate compounds' MoAs to optimize drug development in HD and other diseases.

2.5 Methods

EXPERIMENTAL MODEL AND SUBJECT DETAILS

STHdh Cell Lines

Conditionally immortalized wild-type STHdh^{Q7} (female, Coriell CH00097, RRID: CVCL_M590) and mutant huntingtin homozygous knock-in STHdh^{Q111} (female, Coriell CH00095, RRID: CVCL_M591) murine striatal progenitor cell lines were purchased from Coriell. Cells were maintained at 33°C with 5% CO_2 and cultured in Dulbecco's modified Eagle's medium (DMEM, Corning 10-013) supplemented with 10% fetal bovine serum (FBS, Gemini Bio-Products 100-106), and 1% penicillin/streptomycin (Gemini Bio-Products 400-109).

SH-SY5Y Cell Line

Human neuroblastoma SH-SY5Y (ATCC® CRL-2266™, female, RRID: CVCL_0019) cells were purchased from ATCC. Cells were maintained at 37°C with 5% CO_2 and cultured in a 1:1 mixture of ATCC-formulated Eagle's Minimum Essential Medium (ATCC 30-2003) and F12 medium (ThermoFisher Scientific 11765-054) supplemented with 10% fetal bovine serum (Gemini Bio-Products 100-106).

HEK293T/17 Cell Line

Human embryonic kidney HEK293T/17 (ATCC® CRL-11268™, female, RRID: CVCL_1926, referred to as HEK293 in text) cells were purchased from ATCC. Cells

were maintained at 37°C with 5% CO_2 and cultured in Dulbecco's modified Eagle's medium (DMEM, Corning 10-013) supplemented with 10% fetal bovine serum (FBS, Gemini Bio-Products 100-106), 1% penicillin/streptomycin (Gemini Bio-Products 400-109), and L-glutamine (Sigma-Aldrich G7513).

METHOD DETAILS

Compound Treatment

STHdh cells were incubated in serum-free medium with a compound or vehicle control (DMSO, Sigma-Aldrich 67-68-5) for 24 hours. We chose a treatment time of 24 hours because of the time required to produce a significant cell death phenotype in the $STHdh^{Q111}$ cells. SH-SY5Y and HEK293 cells were incubated in their respective complete medium with a compound or vehicle control (DMSO, Sigma-Aldrich 67-68-5) for 24 hours. The compounds were dissolved in DMSO or water before being added to each medium. For some of the autophagy western blot samples, we also treated the cells for 2 hours with 100nM bafilomycin A1 (Sigma-Aldrich B1793).

Viability Assay

Cell viability was measured using high-content imaging. $STHdh^{Q111}$ cells were seeded at 6,000 cells/well in black 96-well microplates. After 24 hours, the cells were treated with a compound or vehicle. After another 24 hours, 1ug/mL calcein-AM (ThermoFisher Scientific C3099), 2ug/mL propidium iodide (PI, ThermoFisher Scientific P3566), and 1.5ug/mL Hoechst 33442 (ThermoFisher Scientific H3570) were added to detect and quantify live, dead and total cells, respectively. After a 20-minute incubation, the Cellomics Arrayscan Platform (ThermoFisher Scientific) was used for image acquisition and quantitative analysis. ImageJ was used to create composite images (Schneider et al., 2012). $STHdh^{Q7}$ cells with vehicle were also tested using the same procedure, but with a seeding density of 4,500 cells/well to account for the differences in growth rate between the cell lines.

RNA-Seq

RNA was extracted from compound- or vehicle-treated cells in triplicate using Zymo Research Quick-RNA™ MiniPrep (Plus) kit (Zymo Research R1058) and RIN values were tested using Advanced Analytical. All samples had RIN values greater than 0.85. Libraries were prepared using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England Biolabs E7420L) and NEBNext® Poly(A) mRNA Magnetic Isolation Module kit (New England Biolabs E7490L). Libraries were multiplexed and sequenced on an Illumina Hi-Seq 2000 for single-end 50bp reads.

Untargeted Metabolomics

STHdh^{Q111} cells were grown on 10cm dishes in triplicate at a seeding density of 1.06 million cells/well. Compound- or vehicle-treated cells were washed with cold 0.9% NaCl. To each 10cm dish of cells, 660uL LC/MS-grade methanol containing internal standards and 330uL LC/MS-grade water were added. Cells were scraped and transferred to Eppendorf tubes, where 450uL chloroform was added. Samples were vortexed at maximum speed (20,817 rcf) for 10 minutes at 4°C. Each layer was collected separately, avoiding the precipitate at the interface of the two layers, and dried by speedvac. Lipid and polar metabolite profiling were performed by members of the Whitehead Institute Metabolite Profiling Core Facility.

Lipid Profiling

For lipid profiling, cells were resuspended in 50uL 60/35/5 acetonitrile/isopropanol/water (v/v/v) and 5uL was injected for LC/MS analysis. Please see Keckesova et al. and Smulan et al. for a detailed description of the LC/MS analysis (Keckesova et al., 2017; Smulan et al., 2016). Lipid identification and relative quantification was performed using LipidSearch (ThermoFisher Scientific / Mitsui Knowledge Industries). The identified lipids were subjected to quality control filtering and normalization by total signal (Keckesova et al., 2017).

Polar Metabolite Profiling

For polar metabolite profiling, cells were resuspended in 100uL water and 2uL was injected for LC/MS analysis. Please see Birsoy et al. and Chen at al. for a detailed description of the LC/MS analysis (Birsoy et al., 2015; Chen et al., 2016). Untargeted analysis was performed using Progenesis CoMet (Nonlinear Dynamics) using the default settings. Features were filtered based on replicate injections and a dilution series of a pooled sample prepared by mixing equal aliquots of the biological samples. Specifically, the filtering criteria were CV < 0.4 across the four replicate injections and R > 0.9 across a four-point dilution series (comprising 0.1X, 0.3X and 1X concentrations, and a double-volume injection). Features that were not lowest according to the Progenesis quantification in the blank water injection samples were discarded.

H3K4me3 ChIP-Seq

Compound- or vehicle-treated cells were crosslinked with 1% formaldehyde for 8 minutes and quenched with glycine for 5 minutes, lysed in 2X lysis buffer (50mM Tris-HCl pH8, 150mM NaCl, 1% Triton X-100, 0.1% Na Deoxycholate, 5mM CaCl2 and protease inhibitors) for 20 minutes on ice, and digested with 100u MNase (New England Biolabs M0247) for 10 minutes at 37°C. The MNase digestion was terminated by addition of 10mM EDTA. Chromatin was incubated with the anti-H3K4me3 antibody (Millipore 07-473, RRID: AB_1977252) overnight at 4°C, followed by incubation with Protein G beads (Invitrogen 10004D) for 2 hours at 4°C. The beads were washed with PBS (6x) and samples were eluted in EB (10mM Tris-HCl pH8, 5mM EDTA, 300mM NaCl, 0.1% SDS) supplemented with Proteinase K (New England Biolabs P8107S). SPRI beads were used for clean-up and yield was measured using Qubit Fluorimeter. Libraries were prepared using NEBNExt® Ultra™ II DNA Library Prep Kit for Illumina (New England Biolabs E7645S). Libraries were sequenced on an Illumina Hi-Seq 2000 for single-end 50bp reads.

Proteomics

Proteomics were performed by members of the Thermo Fisher Scientific Center for Multiplexed Proteomics at Harvard Medical School. Proteomic data was collected

from cells treated with Group A compounds, Group B compounds, or vehicle controls in triplicate. Please see Weekes et al., McAlister et al., and below for detailed descriptions of the assay (McAlister et al., 2014; Weekes et al., 2014). In brief, sample processing steps included cell lysis, tandem protein digestion using LysC and trypsin, peptide labeling with Tandem Mass Tag 6-plex reagents, IMAC enrichment of phosphopeptides, and peptide fractionation. Multiplexed quantitative mass spectrometry data were collected on an Orbitrap Fusion or Lumos mass spectrometer operating in an MS3 mode using synchronous precursor selection for the MS2 to MS3 fragmentation. Using the SEQUEST algorithm, MS/MS data were searched against a Uniprot mouse database with both the forward and reverse sequences. Additional data processing steps included controlling peptide and protein level false discovery rates, assembling proteins from peptides, and protein quantification from peptides.

Sample Preparation

All solutions are reported as final concentrations. Lysis buffer (8M Urea, 200mM EPPS pH8, Protease and Phosphatase inhibitors from Roche) was added to the vehicle and compound-treated cell pellets. The pellets were vortexed and sonicated to complete cell lysis. Protein concentration of the lysate was determined by micro-BCA assay (Pierce). Proteins were reduced with 5mM TCEP at room temperature for 15 minutes and alkylated with 10mM Iodoacetamide at room temperature for 30 minutes in the dark. Proteins were precipitated using methanol and chloroform. Four volumes of methanol were added to the cell lysate, followed by one volume of chloroform, and three volumes of water. The mixture was vortexed and centrifuged to separate the chloroform phase from the aqueous phase. The precipitated protein was washed with one volume of ice-cold methanol. The washed precipitated protein was air dried. Precipitated protein was resuspended in 200mM EPPS pH8. Proteins were digested with LysC (1:50; enzyme:protein) overnight at room temperature and then further digested with trypsin (1:100; enzyme:protein) for another 8 hours at 37°C. Peptide concentration was quantified using the micro-BCA assay (Pierce). Peptide (100ug) from each condition was labeled with tandem mass tag (TMT6) reagents (1:4; peptide:TMT label) (Pierce) for 2 hours at room temperature. Modification of tyrosine residues with TMT was

reversed by the addition of 5% hydroxylamine for 15 minutes at room temperature. The reaction was quenched with 0.5% TFA. Samples were combined at a 1:1:1:1:1:1 ratio, desalted by C18 solid-phase extraction (SPE, Sep-Pak, Waters), and dried by speedvac.

Phosphopeptide Enrichment

Phosphopeptides were enriched using the High-SelectTM Fe-NTA Phosphopeptide Enrichment Kit (ThermoFisher Scientific). Briefly, the combined TMT6 labeled peptides were resuspended in 200µl binding buffer and incubated with equilibrated resin for 30 minutes at room temperature. Unbound peptides were removed and saved for total protein analysis. Resin was washed and bound peptides were eluted with elution buffer. Eluted peptides were dried by speedvac, resuspended in 1% TFA, desalted by C18 SPE, and dried again. Peptides were resuspended and eluted into glass MS vials from a stage tip packed in-house with 3M Empore resin into two fractions at 20% and 70% ACN with 0.1% Formic acid. Eluted peptides were dried and resuspended in 5% Formic Acid, 5% ACN for MS analysis.

Peptide Fractionation

Peptide fractionation was performed by HPLC bRP. The unbound fraction from IMAC enrichment was dried by speedvac, resuspended in 1% TFA, and cleaned by C18 SPE. The desalted sample was dried by speedvac, resuspended in 5% ACN, 10mM ammonium bicarbonate pH8, and fractionated off-line by basic pH reversed-phase into 96 fractions. Separation was performed using a 50-minute linear gradient from 15% to 45% acetonitrile in 10mM ammonium bicarbonate pH8 at a flow rate of 0.4mL/min over a 300 Extend C18 column (Agilent). Fractions were combined in checkerboard fashion into 24 samples and dried by speedvac.

Liquid Chromatography-MS3 Spectrometry

Of the 24 final fractions from the basic reverse phase, 12 fractions were analyzed with LC-MS3 on an Orbitrap Fusion mass spectrometer (ThermoFisher Scientific) equipped with a Proxeon Easy nLC 1000 for online sample handling and peptide

separations. Approximately 5µg of peptide resuspended in 5% formic acid with 5% acetonitrile was loaded onto a 100µm inner diameter fused-silica micro capillary with a needle tip pulled to an internal diameter less than 5µm. The column was packed in-house to a length of 35cm with a C18 reverse phase resin (GP118 resin 1.8µm, 120Å, Sepax Technologies). The peptides were separated using a 180-minute linear gradient from 3% to 25% buffer B (100% ACN + 0.125% formic acid) equilibrated with buffer A (3% ACN + 0.125% formic acid) at a flow rate of 600nL/min across the column. The scan sequence for the Fusion Orbitrap began with an MS1 spectrum (Orbitrap analysis, resolution 120,000, 350−1,500 m/z scan range, AGC target 4 × 105, maximum injection time 50ms, dynamic exclusion of 120 seconds). The "Top10" precursors were selected for MS2 analysis, which consisted of CID (quadrupole isolation set at 0.7 Da) and ion trap analysis, AGC 1 × 104, NCE 35, maximum injection time 120ms). The top ten precursors from each MS2 scan were selected for MS3 analysis (synchronous precursor selection), in which precursors were fragmented by HCD prior to Orbitrap analysis (NCE 65, max AGC 1 × 105, maximum injection time 150ms, isolation window 2 Da, resolution 50,000).

Phosphopeptide Data Collection

Phosphopeptide samples were analyzed with LC-MS3 on an Orbitrap Lumos mass spectrometer (ThermoFisher Scientific) equipped with a Proxeon Easy nLC 1000 for online sample handling and peptide separations. Total peptide was resuspended in 5% formic acid + 5% acetonitrile was loaded onto a 100µm inner diameter fused-silica micro capillary with a needle tip pulled to an internal diameter less than 5µm. The column was packed in-house to a length of 35cm with a C18 reverse phase resin (GP118 resin 1.8µm, 120Å, Sepax Technologies). The peptides were separated using a 180-minute linear gradient from 3% to 25% buffer B (100% ACN + 0.125% formic acid) equilibrated with buffer A (3% ACN + 0.125% formic acid) at a flow rate of 600nL/min across the column. The scan sequence for the Fusion Orbitrap began with an MS1 spectrum (Orbitrap analysis, resolution 120,000, 400−1,400 m/z scan range, AGC target 1 × 106, maximum injection time 100ms, dynamic exclusion of 120 seconds). The "Top10" precursors were selected for MS2 analysis, which consisted of CID (quadrupole

isolation set at 0.5 Da) and ion trap analysis, AGC 2 × 104, NCE 35, maximum injection time 60ms). The top ten precursors from each MS2 scan were selected for MS3 analysis (synchronous precursor selection), in which precursors were fragmented by HCD prior to Orbitrap analysis (NCE 65, max AGC 2 × 105, maximum injection time 300ms, isolation window 2 Da, resolution 50,000).

LC-MS3 Data Processing and Analysis

A suite of in-house software tools was used for .RAW file processing and controlling peptide and protein level false discovery rates, assembling proteins from peptides, and protein quantification from peptides (McAlister et al., 2014; Weekes et al., 2014). MS/MS spectra were searched against a Uniprot mouse database with both the forward and reverse sequences. Database search criteria are as follows: tryptic with two missed cleavages, a precursor mass tolerance of 50ppm, fragment ion mass tolerance of 1.0 Da, static alkylation of cysteine (57.02146 Da), static TMT labeling of lysine residues and N-termini of peptides (229.162932 Da), and variable oxidation of methionine (15.99491 Da). TMT reporter ion intensities were measured using a 0.003 Da window around the theoretical m/z for each reporter ion in the MS3 scan. Peptide spectral matches with poor quality MS3 spectra were excluded from quantitation (< 100 summed signal-to-noise across 6 channels and < 0.5 precursor isolation specificity). Phosphopeptide searches included variable phosphorylation on serine, threonine, and tyrosine residues (79.96633 Da). Phosphorylation site localization was scored with ModScore. Phosphorylation sites with summed signal-to-noise < 100 across all 6 channels and/or < 0.5 precursor isolation specificity were excluded from quantitation.

Network Analysis

Differential proteins, phosphosites, m/z lipid and polar metabolite peaks, and predicted transcription factors for each compound treatment compared to vehicle control were mapped onto the interactome, comprised of physical interactions between proteins (iRefIndex v14), proteins and metabolites (HMDBv4.0, Recon3D), phosphosites and kinases (PhosphositePlus), m/z peaks and matched metabolites (PIUMet), and phosphosites and proteins (Brunk et al., 2018; Hornbeck et al., 2015; Pirhaji et al.,

2016; Razick et al., 2008; Wishart et al., 2018b). The Prize-Collecting Steiner Forest (PCSF) algorithm was applied using Omics Integrator 2 to find the set of highly relevant pathways associated with each compound treatment (Tuncbag et al., 2016). PCSF was run 100 times with random noise on the edges for robustness measurements and random input sets for specificity measurements. The optimal network solution was filtered by those nodes with at least 40% robustness and specificity.

Autophagic Vacuole Fluorescence Staining

Compound-treated and untreated STHdh^{Q111} and STHdh^{Q7} cells were seeded at 6,000 cells/well in black 96-well microplates. After 24 hours, compounds or vehicle controls were added. After a further 24 hours, the Autophagy Detection Kit (Abcam ab139484) was used to measure autophagic vacuoles in living cells, according to the manufacturer's instructions. Hoechst 33442 (ThermoFisher Scientific H3570) was used to stain the nuclei of cells. Cells with activated autophagy had bright green fluorescent signal. ImageJ was used to create composite images (Schneider et al., 2012). Assay conditions for the SH-SY5Y and HEK293 cells were similar, but with initial seeding concentrations of 25,000 and 5,000 cells/well, respectively.

Western Blots

To quantify LC3 protein expression, adherent cells were scraped in 200µl ice-cold RIPA buffer (50mM Tris-HCl pH8.0, 150mM NaCl, 1% Triton X-100, 0.5% Sodium Deoxycholate, 0.1% SDS supplemented with freshly made protease inhibitors (cOmplete™, EDTA-free Protease Inhibitor Cocktail, Sigma-Aldrich 11873580001)). Samples were incubated with agitation for 30 minutes at 4°C and centrifuged at 12,000×g for 20 minutes at 4°C. The supernatant, containing the protein extracts, was collected. Protein concentration was measured with the Bradford Assay. Protein lysates were separated using SDS/PAGE electrophoresis and transferred to a PVDF membrane. The membranes were rinsed and blocked for 1 hour at room temperature and incubated overnight with primary antibodies in blocking solution with 0.1% Tween-20. The following primary antibodies were used: anti-LC3B (Sigma-Aldrich L7543, dilution 1:500, RRID: AB_796155); anti-Actin (Abcam 1801, dilution 1:1,000). The

membranes were washed and incubated at room temperature for 1 hour with a secondary antibody in a 1:1 PBS, blocking buffer solution with 0.1% Tween-20. The following secondary antibody was used: 800CW Donkey anti-Rabbit IgG (Li-Cor Biosciences 925-32213, dilution 1:10,000, RRID: AB_2715510). The membranes were rinsed and scanned using the Odyssey infrared imaging system (Li-Cor Biosciences). Protein expression was measured using integrated intensity readings in regions around protein bands.

ATP Production Rate Assay

Compound-treated and untreated STHdh^{Q111} and STHdh^{Q7} cells were seeded at 6,000 cells/well in black 96-well microplates. After 24 hours, compounds or vehicle controls were added. After a further 24 hours, the Agilent Seahorse XF Real-Time ATP Rate Assay Kit (Agilent 103592-100) was used to simultaneously measure the rate of ATP productions from mitochondrial respiration and glycolysis, according to manufacturer's instructions. Measurements were taken with the Agilent Seahorse XFe96 analyzer at 33°C. Assay conditions for the SH-SY5Y and HEK293 cells were similar, but with a temperature of 37°C and initial seeding concentrations of 25,000 and 5,000 cells/well, respectively.

QUANTIFICATION AND STATISTICAL ANALYSIS

Protectiveness of Compounds

From the fluorescent images of labeled cells, cell death was quantified as the ratio of PI-positive cells to Hoechst-positive cells using CellProfiler (Carpenter et al., 2006). At least three independent 96-well plates with ten replicate wells each were conducted for each compound and multiple concentrations spanning at least three orders of magnitude were tested. The concentration at which there is minimal cell death is reported for each compound (Table 2-S1). For each experiment, a Student's t-test was applied, and Fisher's method was used to combine the independent experiments and determine significance with a p-value threshold of 0.001. A protective compound in the STHdh^{Q111} model is defined as one that significantly decreased the amount of cell death compared to STHdh^{Q111} vehicle control.

Differentially Expressed Genes

Adapter sequences were trimmed from sequencing reads using Trimmomatic-0.36 (Bolger et al., 2014). Reads were aligned to the GRCm38.p5 transcriptome (https://www.gencodegenes.org/mouse/release_M12.html) and quantified using RSEM (Li and Dewey, 2011). DESeq2 with batch effect modeling by collection day and time was used to find differentially expressed genes for each compound treatment compared to STHdh^{Q111} vehicle control (Love et al., 2014). The differentially expressed genes were filtered using a Benjamini-Hochberg corrected p-value threshold of 0.05.

Differentially Expressed Metabolites

Metabolite quantification in positive and negative ionization mode was log2 normalized and analyzed using limma with batch effect modeling by collection day, and differentially expressed metabolites were filtered using a Benjamini-Hochberg corrected p-value threshold of 0.05 (Ritchie et al., 2015). Untargeted metabolite m/z peaks were matched to known metabolites using PIUMet, with a metabolite database compiled using HMDBv4.0 and Recon3D (Brunk et al., 2018; Pirhaji et al., 2016; Wishart et al., 2018b).

Transcription Factor Prediction

ChIP-Seq adapter sequences were trimmed from sequencing reads using Trimmomatic-0.36 and reads were aligned to the mm10 genome using Bowtie2 (Bolger et al., 2014; Langmead and Salzberg, 2012). Reads were sorted and indexed, and mitochondrial DNA was removed using samtools-1.3 (Li et al., 2009). Peaks were called using MACS2 (Zhang et al., 2008). Motif analysis was used to predict transcription factors that could be regulating the differentially expressed genes. Motifs were annotated to the mm10 UCSC reference genome (http://genome.ucsc.edu/) using the CIS-BP database (Waterston et al., 2002; Weirauch et al., 2014). A hypergeometric test was used for each transcription factor to find those with motifs in regions intersecting ChIP-Seq peaks and within 2kb of differentially expression genes for a given condition. A Benjamini-Hochberg corrected p-value threshold of 0.05 was applied to assign significance to transcription factor predictions.

Differentially Expressed Proteins

Phosphosite quantification was normalized to protein quantification, and both protein and phosphosite data were then log2 normalized and analyzed using limma (Ritchie et al., 2015). Differentially expressed proteins and phosphosites were filtered using a Benjamini-Hochberg corrected p-value threshold of 0.05.

Pathway Enrichment

Enrichment analyses of the differential genes, differential proteins, and network proteins were performed using GOrilla with a background set of all genes measured, all proteins measured, or all proteins present in the interactome, respectively (Eden et al., 2009). Enrichment analyses of the differential metabolites were performed using IMPaLA with a background set of all metabolites measured (Kamburov et al., 2011).

t-SNE Analysis

t-SNE was used to display the transcriptomic and metabolomic data as two-dimensional projections. The inputs were matrices including gene, lipid, or polar metabolite quantifications for each sample and perplexities were set to 15, 14, and 14, respectively. t-SNE analysis was performed in R using the Rtsne package (Krijthe, 2015).

PCA Analysis

Because the number of samples in the proteomic data was lower than in the other omics data types, t-SNE analysis was not applicable. Instead, we displayed the protein and phosphosite data as three-dimensional PCA plots using the stats and rgl packages in R (Adler et al., 2003; R Core Team, 2017).

Network Visualization

Networks were visualized in Cytoscape (Shannon et al., 2003). In each network, the nodes are proteins, phosphosites, transcription factors, or metabolites. The proteomic data are mapped onto proteins and phosphosites. The integration of the RNA-Seq and ChIP-Seq data provided transcription factor predictions. The metabolite

data is shown as metabolite peaks connected to m/z-matched known metabolites. The edges represent the physical interactions between the molecules. Bigger nodes are more robust, as determined by the PCSF randomizations. The red and blue colors indicate the log2 fold change, as determined by the omics data.

Western Blot Analysis

The LC3-II/LC3-I ratio was calculated for each sample, with and without the addition of bafilomycin A1 (bafA). For each compound condition without bafA, a Student's t-test was applied to the three replicates of the compound-treated samples compared to the vehicle control samples. Significance was determined with a p-value threshold of 0.05. For each compound condition with bafA, a Student's t-test was applied to the three replicates of the bafA-treated samples compared to their respective condition's bafA-untreated samples. Significance was determined with a p-value threshold of 0.1. The LC3-II/LC3-I ratios for samples in each cell line are normalized to their respective controls in the quantification, such that the control samples are have a ratio of 1 (Figures 2-6, 2-7).

Quantifying ATP Production Rates

For each compound condition, a Student's t-test was applied to the data for the replicates (at least 6 per treatment) of the compound-treated samples compared to the vehicle control samples.

Dendrogram Clustering

Using the controls and the four compounds in Groups A and B analyzed with all of the omics data, a distance matrix was calculated for each data type using the Euclidean distance measure in the stats package in R (R Core Team, 2017). Dendrograms were created using the distance matrices for each data type using the hclust function with the Ward clustering method in the stats package in R (R Core Team, 2017).

Calculating Structural Similarities

Identifiers for each compound were uploaded to ChemMine tools and the "Similarity Workbench" feature was used to compare each pair of compounds. The tool calculates atom pair and reports maximum common substructure (MCS) scores with the Tanimoto coefficient as the similarity measure (Backman et al., 2011).

Calculating Connectivity Similarities

The L1000 connectivity scores between pairs of compounds were assessed using the "Touchstone" analysis tool as part of the Connectivity Map (Lamb et al., 2006; Subramanian et al., 2017). Only eight of the 30 compounds profiled were part of the Connectivity Map dataset.

DATA AND SOFTWARE AVAILABILITY

Deposited Data

The RNA-Seq and ChIP-Seq data have been deposited in the Gene Expression Omnibus with accession number GSE129144.

2.6 Supplemental Information

Supplemental Information includes four figures and eight tables.

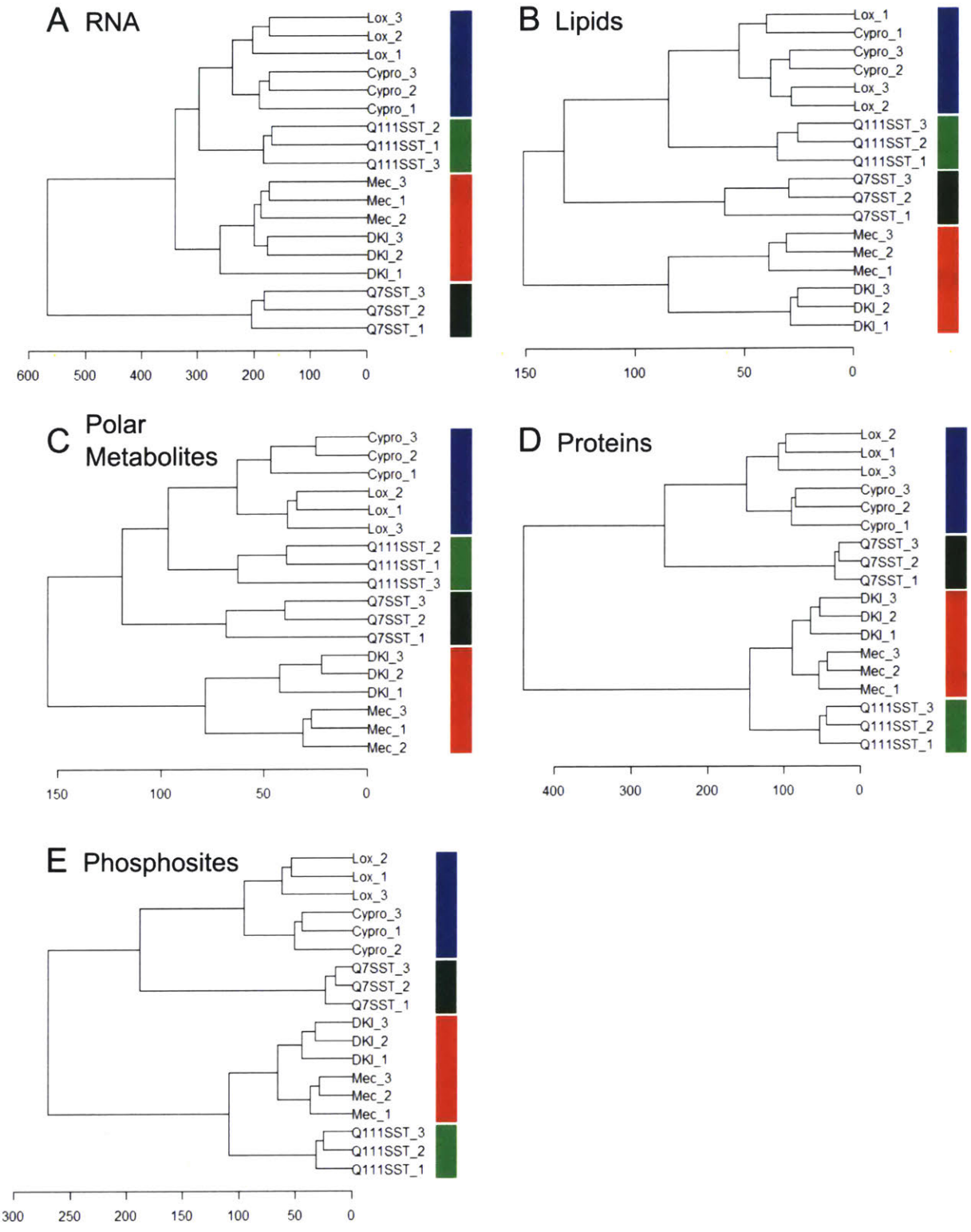

Figure 2-S1 Distinct omics data lead to different clustering patterns between the compound-treated and control samples.

Related to Figure 2-4.

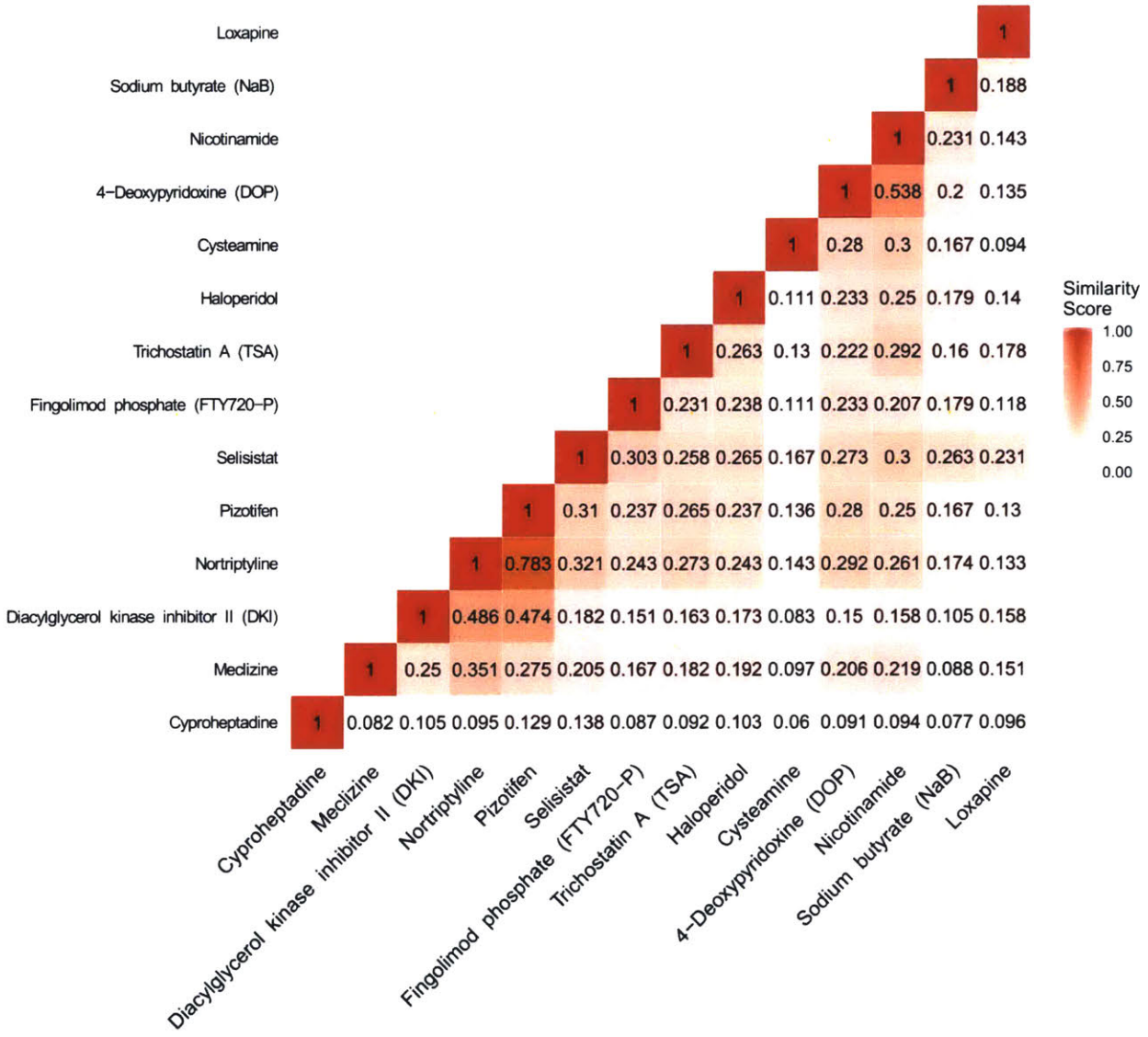

Figure 2-S2 Groupings of compounds would not be predicted based on structural similarities determined by maximum common substructure (MCS) Tanimoto coefficients.

Each number within the matrix indicates the similarity score calculated using the MCS Tanimoto coefficients. Cells on the diagonal were assigned a similarity score of 1. Related to Figure 2-4.

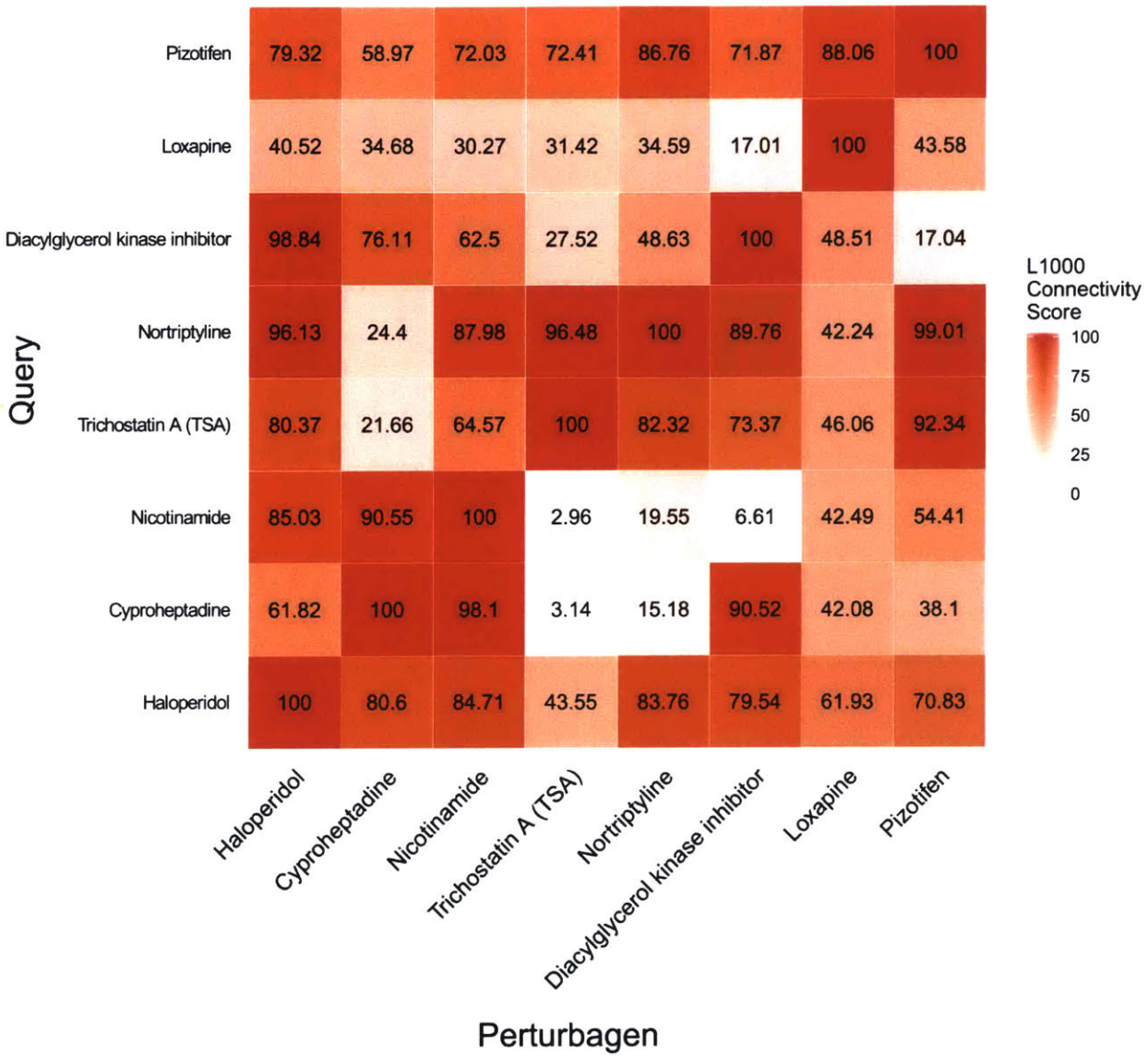

Figure 2-S3 Groupings of compounds would not be predicted based on their L1000 connectivity scores.

Each number within the matrix indicates the connectivity score, where the compound on the y-axis was used as the query and the compound on the x-axis was used as the perturbagen. Cells on the diagonal were assigned a connectivity score of 100. Related to Figure 2-4.

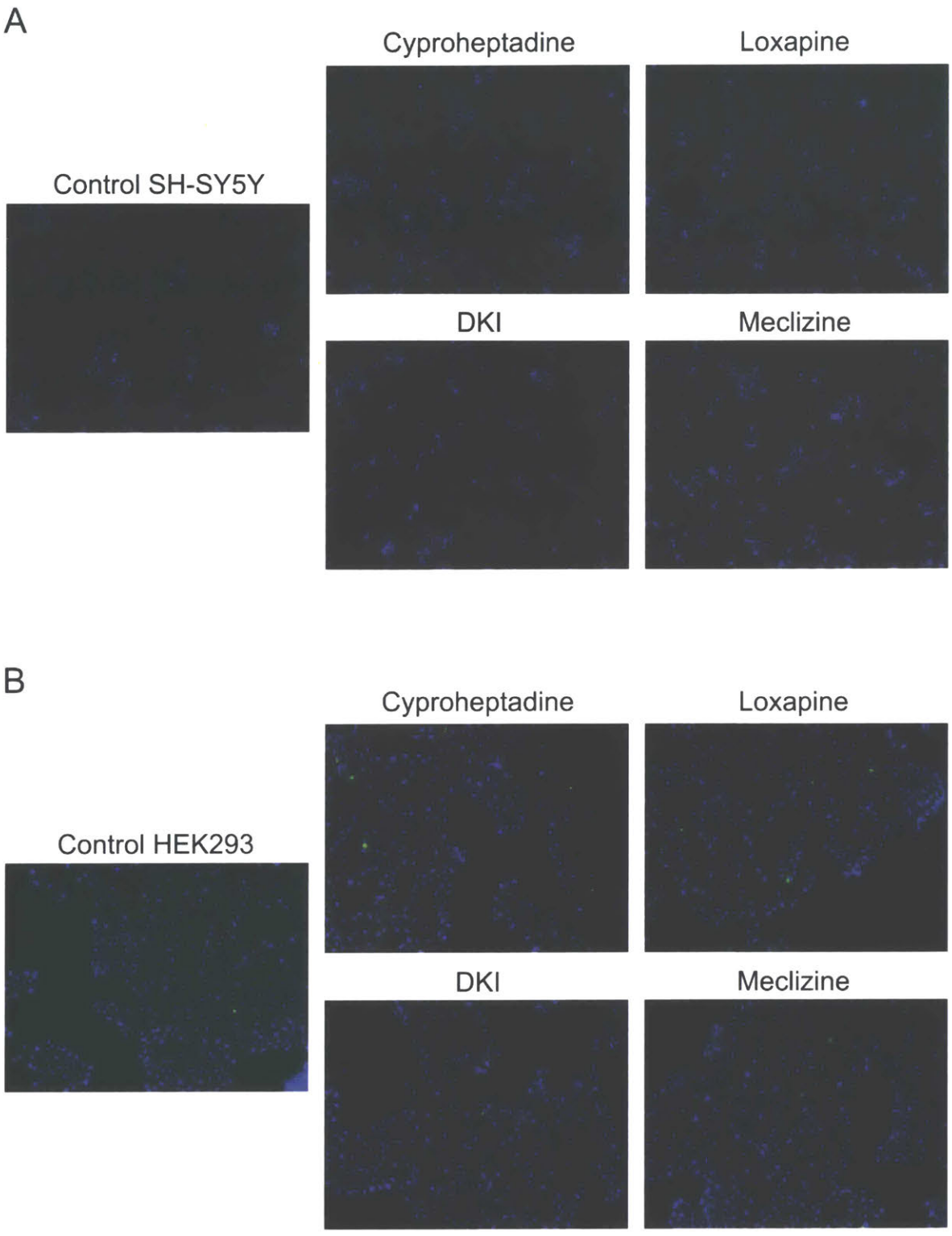

Figure 2-S4 Staining of autophagic vacuoles is increased by Group A compounds in (A) SH-SY5Y and (B) HEK293 cells.

Related to Figure 2-7.

Table 2-S1 Dose, vendor, literature reference, FDA-approval status, and known targets for the 30 tested compounds.

Related to Figure 2-3.

Compound	Dose	Vendor	Literature Reference	FDA Approval	Known DrugBank Targets
Meclizine	10uM	Sigma-Aldrich	Gohil et al., 2013	Yes	HRH1, NR1I3
Sodium butyrate	1mM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	No	--
Cyproheptadine	10uM	Sigma-Aldrich	Sarantos, Papanikolaou, Ellerby, and Hughes, 2012	Yes	HRH1, HTR2A, HTR2C, CHRM1, CHRM2, CHRM3, HTR7
Loxapine	10uM	Sigma-Aldrich	Sarantos, Papanikolaou, Ellerby, and Hughes, 2012	Yes	DRD2, DRD1, HTR2A, HTR2C, HTR1A, HTR1B, HTR1D, HTR1E, HTR3A, HTR5A, HTR6, HTR7, ADRA1A, ADRA1B, ADRA2A, ADRA2B, ADRA2C, ADRB1, CHRM1, CHRM2, CHRM3, CHRM4, CHRM5, DRD3, DRD4, DRD5, HRH1, HRH2, HRH4, SLC6A4, SLC6A2, SLC6A3
4-Deoxypyridoxine	4mM	Sigma-Aldrich	Pirhaji et al., 2017	No	--
Selisistat	10uM	Selleckchem	Westerberg et al., 2015	No	SIRT1
Trichostatin A	10nM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	No	--
Diacylglycerol kinase inhibitor II	10uM	Sigma-Aldrich	Zhang et al., 2012	No	--
Nicotinamide	0.5nM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	ETA, LDHA, PARP1, SIRT5, BST1
Nortriptyline	1nM	Sigma-Aldrich	Lauterbach et al., 2013	Yes	SLC6A2, SLC6A4, HTR2A,

					HTR1A, HRH1, ADRA1A, ADRA1D, CHRM1, CHRM2, CHRM3, CHRM4, CHRM5, HTR2C, HTR6, ADRA1B, DRD2
Fingolimod phosphate	250nM	Santa Cruz Biotechnology	Pirhaji et al., 2016	No	--
Haloperidol	0.5nM	Cayman Chemical	Lauterbach et al., 2013	Yes	DRD2, DRD1, GRIN2B, HTR2A, DRD3, MCHR1, SLC18A2
Pizotifen	5uM	Sigma-Aldrich	Sarantos, Papanikolaou, Ellerby, and Hughes, 2012	Yes	CHRM1, CHRM2, CHRM3, HTR2A, HTR2B, HTR2C, HTR1A, HTR1B, HTR1D, HRH1, ADRA1A, ADRA1B, ADRA1D, ADRA2A, ADRA2B, ADRA2C
Cysteamine	250uM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	Cystine, SST, NPY2R
Sodium phenylbutyrate	100nM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	TYRB, NPR
Methylene blue	1nM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	GUCY1A2, NOS1
Rapamycin	1nM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	MTOR, FKBP1A, FGF2
Bezafibrate	100nM	Cayman Chemical	Chandra et al., 2016	Yes	PPARA, PPARD, PPARG, NR1I2, RXRA, RXRB, RXRG
(-)-Epigallocatechin gallate	100nM	Cayman Chemical	Zuccato, Valenza, and Cattaneo, 2010	No	AHR, DNMT1, DHFRL1
Creatine	500uM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	CKM, CKMT1A, CKB, CKMT2, SLC6A8, GAMT

Cystamine	250uM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	No	--
Desipramine	1uM	Sigma-Aldrich	Lauterbach et al., 2013	Yes	SLC6A2, SLC6A4, HTR2A, ADRB2, ADRB1, SMPD1, HRH1, ADRA1A, ADRA1B, ADRA1D, CHRM1, CHRM2, CHRM3, CHRM4, CHRM5, HTR1A, HTR2C, DRD2, ADRA2A, ADRA2B, ADRA2C
7,8-Dihydroxyflavone	100nM	Cayman Chemical	Jiang et al., 2013	No	--
Minocycline	10uM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	RPSL, RPSD, IL1B, ALOX5, MMP9, VEGFA, CASP1, CASP3, CYCS
Melatonin	10nM	Sigma-Aldrich	Lauterbach et al., 2013	Yes	MTNR1A, MTNR1B, ESR1, RORB, CALM1, MPO, EPX, CALR, ASMT, NQO2
Suberoylanilide hydroxamic acid	1nM	Sigma-Aldrich	Hockly et al., 2003	Yes	HDAC1, HDAC2, HDAC3, HDAC6, HDAC8, ACUC1
Curcumin	10nM	Sigma-Aldrich	Bates, Tabrizi, and Jones, 2014	Yes	PPARG, VDR, ABCC5, CBR1, GSTP1
Celastrol	1nM	Sigma-Aldrich	Wang, Gines, MacDonald, and Gusella, 2005	No	--
Fingolimod	1uM	Sigma-Aldrich	Di Pardo et al., 2014	Yes	S1PR5, HDAC1
Juglone	1nM	Sigma-Aldrich	Wang, Gines, MacDonald, and Gusella, 2005	No	--

Table 2-S4 GO enrichment for the differentially expressed proteins affected by Group A compounds.

Related to Figure 2-4.

GO Term	Description	P-value	FDR-adjusted p-value	Enrichment score	Number of total proteins (#TotP)	#TotP in GO term	Number of Group A differential proteins (#DEP_A)	#DEP_A in GO term
GO:0030199	Collagen fibril organization	1.38 E-05	4.07E-02	2.48	6098	26	1703	18
GO:0043413	Macro-molecule glycosylation	1.07 E-05	4.21E-02	1.9	6098	68	1703	36

Supplemental Excel tables can be found in the publication:

Table 2-S2 GO enrichment for the differentially expressed genes affected by Group A compounds.
Related to Figure 2-4.

Table 2-S3 Pathway enrichment using IMPaLA for the differentially expressed metabolites affected by Group A compounds.
Related to Figure 2-4.

Table 2-S5 GO enrichment for the differentially expressed genes affected by Group B compounds.
Related to Figure 2-4.

Table 2-S6 Pathway enrichment using IMPaLA for the differentially expressed metabolites affected by Group B compounds.
Related to Figure 2-4.

Table 2-S7 GO enrichment for the proteins in the Group A network.
Related to Figure 2-5.

Table 2-S8 GO enrichment for the proteins in the Group B network.
Related to Figure 2-5.

2.7 References

Adler, D., Nenadić, O., and Zucchini, W. (2003). RGL: A R-library for 3D visualization with OpenGL. In Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics.

Backman, T.W.H., Cao, Y., and Girke, T. (2011). ChemMine tools: An online service for analyzing and clustering small molecules. Nucleic Acids Res. *39*, W486-91.

Bates, G., Tabrizi, S.J., and Jones, L. (2014). Huntingtin's Disease (Oxford: Oxford University Press).

Birsoy, K., Wang, T., Chen, W.W., Freinkman, E., Abu-Remaileh, M., and Sabatini, D.M. (2015). An Essential Role of the Mitochondrial Electron Transport Chain in Cell Proliferation Is to Enable Aspartate Synthesis. Cell *162*, 540–551.

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

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. *36*, 272–281.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100.

Chen, W.W., Freinkman, E., Wang, T., Birsoy, K., and Sabatini, D.M. (2016). Absolute Quantification of Matrix Metabolites Reveals the Dynamics of Mitochondrial Metabolism. Cell *166*, 1324–37.e11.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Gaulton, A., Hersey, A., Nowotka, M.L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrian-Uhalte, E., et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res. *45*, D945-54.

Gohil, V.M., Zhu, L., Baker, C.D., Cracan, V., Yaseen, A., Jain, M., Clish, C.B., Brookes, P.S., Bakovic, M., and Mootha, V.K. (2013). Meclizine inhibits mitochondrial respiration through direct targeting of cytosolic phosphoethanolamine metabolism. J. Biol. Chem. *288*, 35387–35395.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Hong, C.T., Chau, K.Y., and Schapira, A.H.V. (2016). Meclizine-induced enhanced glycolysis is neuroprotective in Parkinson disease cell models. Sci. Rep. *6*.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512-20.

Iorio, F., Saez-Rodriguez, J., and Bernardo, D. di (2013). Network based elucidation of drug response: From modulators to targets. BMC Syst. Biol. *13*, 7–139.

Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R., and Keun, H.C. (2011). Integrated

pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics *27*, 2917–2918.
Keckesova, Z., Donaher, J.L., De Cock, J., Freinkman, E., Lingrell, S., Bachovchin, D.A., Bierie, B., Tischler, V., Noske, A., Okondo, M.C., et al. (2017). LACTB is a tumour suppressor that modulates lipid metabolism and cell state. Nature *543*, 681–686.
Kedaigle, A., and Fraenkel, E. (2018). Turning omics data into therapeutic insights. Curr. Opin. Pharmacol. *42*, 95–101.
Kedaigle, A., Fraenkel, E., Atwal, R., Wu, M., Gusella, J., MacDonald, M., Kaye, J., Finkbeiner, S., Mattis, V., Tom, C., et al. (2019). Bioenergetic deficits in Huntington's disease iPSC-derived neural cells and rescue with glycolytic metabolites. Hum. Mol. Genet.
Krijthe, J.H. (2015). T-Distributed Stochastic Neighbor Embedding using Barnes-Hut.
Kumar, A., Kumar Singh, S., Kumar, V., Kumar, D., Agarwal, S., and Rana, M.K. (2015). Huntington's disease: An update of therapeutic strategies. Gene *556*, 91–97.
Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science (80-.). *313*, 1929–1935.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
Litichevskiy, L., Peckner, R., Abelin, J.G., Asiedu, J.K., Creech, A.L., Davis, J.F., Davison, D., Dunning, C.M., Egertson, J.D., Egri, S., et al. (2018). A Library of Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations. Cell Syst. *6*, 424–43.e7.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
Martin, D.D.O., Ladha, S., Ehrnhoefer, D.E., and Hayden, M.R. (2015). Autophagy in Huntington disease and huntingtin in autophagy. Trends Neurosci. *38*, 26–35.
McAlister, G.C., Nusinow, D.P., Jedrychowski, M.P., Wühr, M., Huttlin, E.L., Erickson, B.K., Rad, R., Haas, W., and Gygi, S.P. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal. Chem. *86*, 7150–7158.
Mizushima, N., and Yoshimori, T. (2007). How to interpret LC3 immunoblotting. Autophagy *3*, 542–545.
Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C.B., White, F.M., Saghatelian, A., and Fraenkel, E. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. Nat. Methods *13*, 770–776.
Pirhaji, L., Milani, P., Dalin, S., Wassie, B.T., Dunn, D.E., Fenster, R.J., Avila-Pacheco, J., Greengard, P., Clish, C.B., Heiman, M., et al. (2017). Identifying therapeutic targets by combining transcriptional data with ordinal clinical measurements. Nat.

Commun. *8*, 623.
R Core Team (2017). R: A language and environment for statistical computing. http://www.R-project.org/.
Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.
Rees, M.G., Seashore-Ludlow, B., Cheah, J.H., Adams, D.J., Price, E. V., Gill, S., Javaid, S., Coletti, M.E., Jones, V.L., Bodycombe, N.E., et al. (2016). Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat. Chem. Biol. *12*, 109–116.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
Sarantos, M.R., Papanikolaou, T., Ellerby, L.M., and Hughes, R.E. (2012). Pizotifen activates ERK and provides neuroprotection in vitro and in vivo in models of Huntington's disease. J. Huntingtons. Dis. *1*, 195–210.
Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.
Schulte, J., and Littleton, J.T. (2011). The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. Curr. Trends Neurol. *5*, 65–78.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.
Smulan, L.J., Ding, W., Freinkman, E., Gujja, S., Edwards, Y.J.K., and Walker, A.K. (2016). Cholesterol-Independent SREBP-1 Maturation Is Linked to ARF1 Inactivation. Cell Rep. *16*, 9–18.
Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437–52.e17.
Trettel, F., Rigamonti, D., Hilditch-Maguire, P., Wheeler, V.C., Sharp, a H., Persichetti, F., Cattaneo, E., and MacDonald, M.E. (2000). Dominant phenotypes produced by the HD mutation in STHdh(Q111) striatal cells. Hum. Mol. Genet. *9*, 2799–2809.
Tulloch, L.B., Menzies, S.K., Coron, R.P., Roberts, M.D., Florence, G.J., and Smith, T.K. (2018). Direct and indirect approaches to identify drug modes of action. IUBMB Life *70*, 9–22.
Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput. Biol. *12*, e1004879.
Varma, H., Lo, D., and Stockwell, B. (2008). High Throughput Screening for Neurodegeneration and Complex Disease Phenotypes. Comb. Chem. High Throughput Screen. *11*, 238–248.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.
Weekes, M.P., Tomasec, P., Huttlin, E.L., Fielding, C.A., Nusinow, D., Stanton, R.J.,

Wang, E.C.Y., Aicheler, R., Murrell, I., Wilkinson, G.W.G., et al. (2014). Quantitative temporal viromics: An approach to investigate host-pathogen interaction. Cell *157*, 1460–1472.

Wehling, M. (2009). Assessing the translatability of drug projects: What needs to be scored to predict success? Nat. Rev. Drug Discov. *8*, 541–546.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell *158*, 1431–1443.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018a). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074-82.

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., et al. (2018b). HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Res. *46*, D1074-82.

Wong, C.H., Siah, K.W., and Lo, A.W. (2018). Estimation of clinical trial success rates and related parameters. Biostatistics *20*, 273–286.

Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Rodríguez Martínez, M., López, G., Mattioli, M., Realubit, R., et al. (2015). Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell *162*, 441–451.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955-61.

Zampieri, M., Szappanos, B., Buchieri, M.V., Trauner, A., Piazza, I., Picotti, P., Gagneux, S., Borrell, S., Gicquel, B., Lelievre, J., et al. (2018). High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. Sci. Transl. Med. *10*.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zuccato, C., Valenza, M., and Cattaneo, E. (2010). Molecular Mechanisms and Potential Therapeutical Targets in Huntington ' s Disease. Physiol Rev *90*, 905–981.

Chapter 3: Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells

This work was published in 2016.

Milani P, Escalante-Chong R, Shelley BC, Patel-Murray NL, Xin X, Adam M, Mandefro B, Sareen D, Svendsen CN, and Fraenkel E (2016). Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. Sci. Rep. *6*.

As part of this work, I would like to acknowledge members of the MIT BioMicro Center for assistance with sequencing data collection. I would also like to acknowledge SMA patients and their families for their essential contributions to this research.

My contributions:
I assisted with the ATAC-Seq protocol optimization and experiments. I also contributed to the writing of the manuscript.

3.1 Abstract

In recent years, the assay for transposase-accessible chromatin using sequencing (ATAC-Seq) has become a fundamental tool of epigenomic research. However, it is difficult to perform this technique on frozen samples because freezing cells before extracting nuclei can impair nuclear integrity and alter chromatin structure, especially in fragile cells such as neurons. Our aim was to develop a protocol for freezing neuronal cells that is compatible with ATAC-Seq; we focused on a disease-relevant cell type, namely motor neurons differentiated from induced pluripotent stem cells (iMNs) from a patient affected by spinal muscular atrophy. We found that while flash-frozen iMNs are not suitable for ATAC-Seq, the assay is successful with slow-cooled cryopreserved cells. Using this method, we were able to isolate high quality, intact nuclei, and we verified that epigenetic results from fresh and cryopreserved iMNs quantitatively agree.

3.2 Introduction

Since its establishment, the assay for transposase-accessible chromatin using sequencing (ATAC-Seq) has revolutionized the study of epigenetics (Buenrostro et al., 2013, 2015a). This technique detects open-chromatin regions and maps transcription factor binding events genome-wide by means of direct in vitro transposition of native chromatin. Specifically, hyperactive Tn5 transposase is used to interrogate chromatin accessibility by inserting high-throughput DNA sequencing adapters into open genomic regions, which allows for the preferential amplification of DNA fragments located at sites of active chromatin. Because the DNA sites directly bound by DNA-binding proteins are protected from transposition, this approach enables the inference of transcription factor occupancy at the level of individual functional regulatory regions. Furthermore, ATAC-Seq can be utilized to decode nucleosome occupancy and positioning, by exploiting the fact that the Tn5 transposase cuts DNA with a periodicity of about 150-200bp, corresponding to the length of the DNA fragments wrapped around histones (Schep et al., 2015). This periodicity is maintained up to six nucleosomes and provides information about the spatial organization of nucleosomes within accessible chromatin. ATAC-Seq signals thus allow for the delineation of fine-scale architectures of the regulatory

framework by correlating occupancy patterns with other features, such as chromatin remodeling and global gene induction programs.

Compared to other epigenetic methodologies, such as FAIRE-Seq and conventional DNase-Seq, ATAC-Seq requires a small number of cells. Therefore, it is suitable for work on precious samples, including differentiated cells derived from induced pluripotent stem cells (iPSCs), primary cell culture, and limited clinical specimens. Recently developed techniques, such as single-cell DNase sequencing (scDNase-seq), indexing-first ChIP-Seq (iChIP), ultra-low-input micrococcal nuclease-based native ChIP (ULI-NChIP), and ChIPmentation, allow for the epigenomic investigation of small number of cells or even single cells without requiring microfluidic devices (Brind'Amour et al., 2015; Jin et al., 2015; Lara-Astiaso et al., 2014; Schmidl et al., 2015). However, these assays require multiple experimental steps. In contrast, in ATAC-Seq the actual assay and library preparation are performed in a single enzymatic reaction. Hence, this technique is less time-consuming and labor-intensive.

It is essential to preserve the native chromatin architecture and the original nucleosome distribution patterns for ATAC-Seq. Freezing samples prior to the purification of nuclei can be detrimental to nuclear integrity and can affect chromatin structures, thus restricting the application of ATAC-Seq to freshly-isolated nuclei (Trusal et al., 1984). This limits the use of ATAC-Seq on clinical samples, which are typically stored frozen, and represents a major logistical hurdle for long-distance collaborative projects, for which sample freezing is often inevitable.

In an attempt to overcome this drawback, we identified a freezing protocol suitable for native chromatin-based assays on neuronal cells. We tested the freezing techniques using a disease-relevant cell type, namely motor neurons (iMNs) differentiated from human iPSCs, which were derived from the fibroblasts of a patient affected by spinal muscular atrophy (SMA). This disease is caused by homozygous loss of the *SMN1* gene and is characterized by the degeneration of lower motor neurons (Ogino and Wilson, 2004).

We tested two different freezing methods: flash-freezing and slow-cooling cryopreservation. Flash-freezing is a procedure in which the temperature of the sample is rapidly lowered using liquid nitrogen, dry ice or dry ice/ethanol slurry, in order to limit

the formation of damaging ice crystals. Conversely, slow-cooling cryopreservation lowers the temperature of the sample gradually and makes use of cryoprotectants, such as dimethyl sulfoxide (DMSO), to prevent ice crystal nucleation and limit cell dehydration during freezing. Cryopreservation techniques are widely employed for cell banking purposes and are routinely used in assisted reproduction technologies (Dovey, 2012; Paramanantham et al., 2015).

We introduced a number of experimental quality control (QC) checkpoints and steps for data analysis to monitor the efficacy of the procedures and quantify potential alterations induced by cell freezing.

3.3 Results and Discussion

3.3.1 Description of experimental design and overview of the protocol

We generated ATAC-Seq data on fresh (F), flash-frozen (FF), and cryopreserved (C) iMNs by following the procedure outlined in Figure 3-1. Fresh and frozen neurons were derived from the same pool of cells and processed in parallel in order to estimate the effects of freezing on ATAC-Seq outcomes without any batch effect bias.

The ATAC-Seq protocol was adapted from Buenrostro et al., with some modifications (Buenrostro et al., 2013, 2015b). Given that a successful ATAC-Seq experiment begins with the isolation of high-quality intact nuclei, we first introduced a quality control checkpoint consisting of the morphological evaluation of nuclei with either Trypan Blue or DAPI staining, followed by the accurate quantification of those nuclei using an automated cell counter. Precise counting of nuclei is important to ensure optimal tagmentation (the simultaneous fragmenting of the DNA and insertion of adapter sequences) and to limit the technical variability across samples. From a qualitative perspective, individual intact nuclei with a round or oval shape should be observed with no visible clumping. To exclude samples with severe degradation or over-tagmentation, we assessed the quality of the treated chromatin samples by gel electrophoresis, as described in Buenrostro et al. (Buenrostro et al., 2015b); if the chromatin was intact and the transposase reaction was optimal, a DNA laddering pattern with a periodicity of about 200bp should be observed, corresponding to fragments of DNA that were originally protected by an integer number of nucleosomes (nucleosome phasing).

Furthermore, we measured the enrichment of DNA accessible regions by performing real-time qPCR analysis using known open-chromatin sites as positive controls and Tn5-insensitive sites as negative controls. When assayed by real-time qPCR, high-quality ATAC-Seq samples should show at least a 10-fold enrichment of positive control sites compared to Tn5-insensitive sites. Finally, as we were principally interested in open-chromatin profiling and not in nucleosome positioning, we introduced a size-selection step to enrich for nucleosome-free fragments. After size-selection, libraries were PCR-amplified and submitted for single-end sequencing.

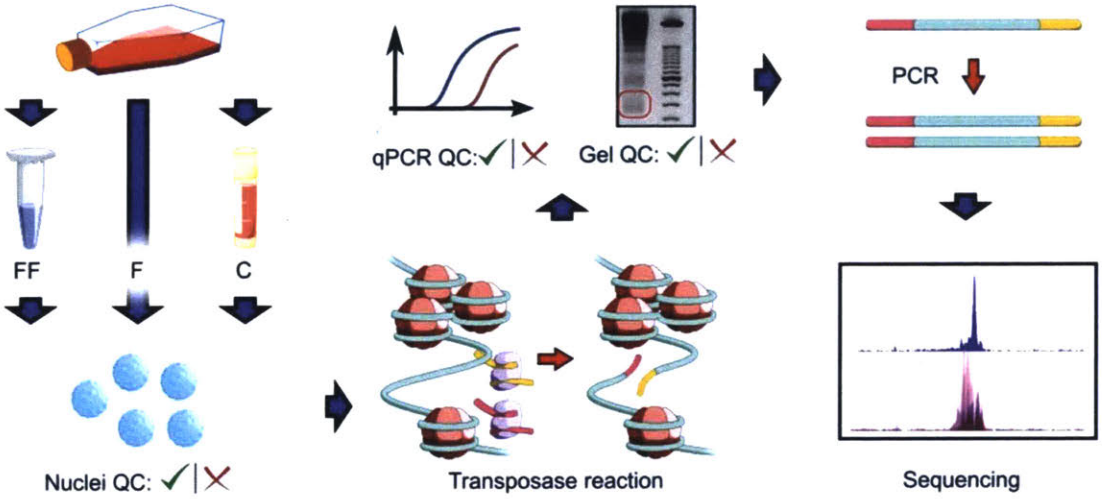

Figure 3-1 Outline of ATAC-Seq procedure using fresh, flash-frozen, and cryopreserved iPSC-derived motor neurons.

The key experimental steps are nuclei extraction, transposase reaction, size selection, PCR amplification and sequencing. The quality control (QC) checkpoints consist of morphological evaluation of nuclei, agarose gel electrophoresis of libraries, and real-time qPCR to assess the enrichment of open-chromatin sites. (F = fresh, FF = flash-frozen, C = cryopreserved).

3.3.2 ATAC-Seq on iPSC-derived motor neurons (iMNs): flash-frozen cells

We first performed ATAC-Seq on fresh and flash-frozen iMNs. Differentiated neuronal cells were generated as described in Methods. We performed

immunocytochemistry experiments using antibodies against markers of mature motor neurons to test the efficiency of the differentiation protocol; we showed that patient-derived iPSCs were successfully differentiated into ISL1- and SMI32-positive motor neurons (Figure 3-2). Figure 3-3 shows ATAC-Seq outcomes from two representative samples. Nuclei from fresh cells passed quality control, while nuclei from flash-frozen neurons exhibited excessive clumping, likely caused by disruption of the nuclear envelope and consequent leakage of DNA (Figure 3-3A). After the transposase reaction, we assessed the quality of the resulting libraries by qualitative evaluation of agarose gel electrophoresis. The library from freshly-isolated nuclei displayed clear nucleosome phasing, while the library from flash-frozen neurons showed DNA smearing on the gel (Figure 3-3B). This result strongly indicates that loss of chromatin integrity occurred during flash-freezing. We proceeded with next-generation sequencing for one fresh and one flash-frozen sample. We used the R package Gviz to plot the sequencing data along genomic coordinates for manual inspection of tracks and local visualization of peaks (Figures 3-3C and 3-S1). As a negative control, we treated human naked DNA with the hyperactive Tn5 enzyme and sequenced this library alongside the ATAC-Seq samples. ATAC-Seq peaks from fresh neurons were sharp and overlapped with H3K4me3 signals from ENCODE ChIP-Seq datasets. Using a MACS2 q-value threshold of 0.05, we obtained more than seventy thousand significant peaks using fresh cells. In contrast, the reads from flash-frozen cells were distributed evenly across the entire genome, similar to the results obtained with the negative control, and only 461 significant peaks were detected. Half of these peaks overlapped with the peaks from fresh iMNs (Figure 3-S2).

These findings indicate that flash-freezing of iMNs is not suitable for ATAC-Seq.

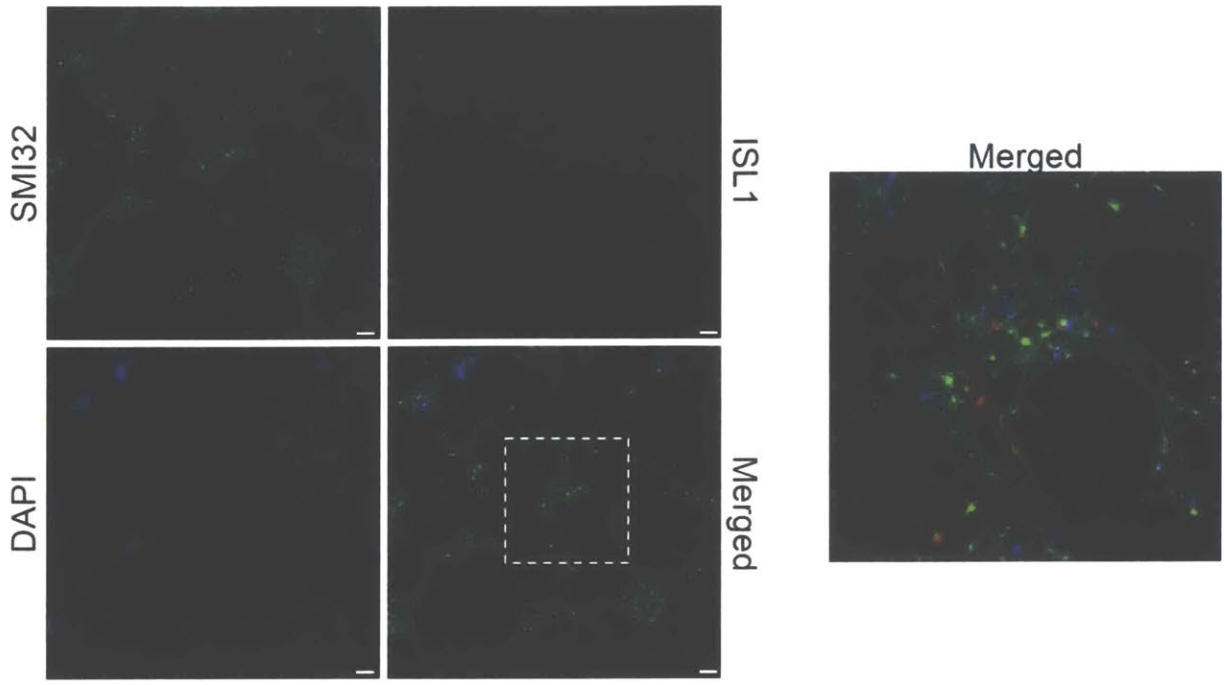

Figure 3-2 Fibroblast-derived iPSCs differentiate into SMI32- and ISL1-positive motor neurons.

Differentiated cells were labeled to evaluate the immunoreactivity of SMI32 (green) and ISL1 (red) proteins, two markers of mature motor neurons. Nuclei were stained with DAPI. Motor neurons were imaged with 10x magnification. The image on the right represents a higher magnification of selected neurons. Scale bar = 75 μm.

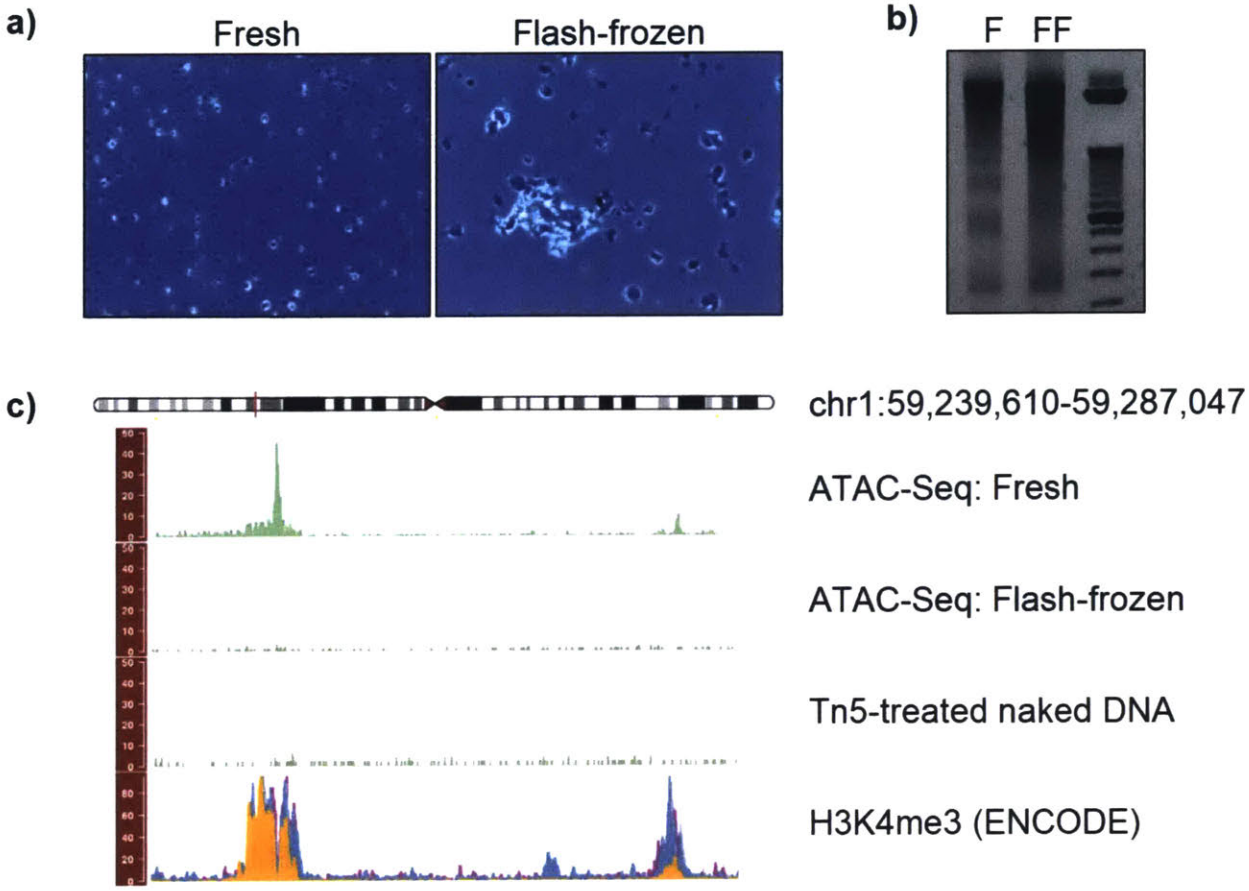

Figure 3-3 Representative results for ATAC-Seq carried out on fresh and flash-frozen cells.

(A) Nuclear morphological evaluation: nuclei from fresh cells were of high quality, while excessive clumping was observed for nuclei from flash-frozen neurons. (B) Agarose gel electrophoresis of libraries: the nucleosome phasing pattern on the gel was not detected in flash-frozen samples, as opposed to fresh cells. (C) ATAC-Seq tracks were visualized with the Gviz package: while we detected sharp peaks for fresh samples, the reads from flash-frozen neurons were distributed noisily across the genome. (F = fresh, FF = flash-frozen).

3.3.3 ATAC-Seq on iPSC-derived motor neurons (iMNs): cryopreserved cells

Next, we compared ATAC-Seq results from fresh and cryopreserved cells. Approximately one million fresh iMNs were transferred to Cryostor media and slowly frozen, stored, and then thawed for processing. After thawing, we assessed the cell death rate by evaluating chromatin condensation, which is a hallmark of apoptotic cells (Ziegler and Groscurth, 2004). To this purpose, we stained the neurons with the cell-permeable Hoechst 33342, then quantified chromatin condensation using fluorescent microscopy. This dye brightly stains the condensed chromatin of cells undergoing apoptosis (Figure 3-S3). The rate of cell death was 10.8% with standard deviation of 1.7; the fraction of nuclei recovered was higher than 70% (Table 3-S1). As shown in Figure 3-4A, nuclei from the cryopreserved cells were of high quality and the nucleosome laddering was detected by gel electrophoresis (Figure 3-4B). Sequencing data from both fresh and cryopreserved samples showed sharp peaks and low background signal (Figures 3-4C and 3-S1). Furthermore, the qPCR enrichment of the positive control site (GAPDH gene promoter, Figure 3-5A top panel) over the Tn5-insensitive site (gene desert region, Figure 3-5A bottom panel) was high and comparable to that of fresh cells, as opposed to qPCR results from flash-frozen neurons, for which less than 10-fold enrichment was observed (Figure 3-5B). We obtained similar results using a second set of primers designed to amplify open-chromatin and gene desert regions (Figure 3-S4). As in the case of fresh cells, we obtained more than seventy thousand significant peaks using cryopreserved samples (MACS2 q-value threshold = 0.05) (Table 3-1). There was high overlap in the number of peaks obtained from fresh and cryopreserved iMNs (Figure 3-S5). These results reveal that slow-cooling cryopreservation of iMNs is compatible with native chromatin-based epigenetic assays.

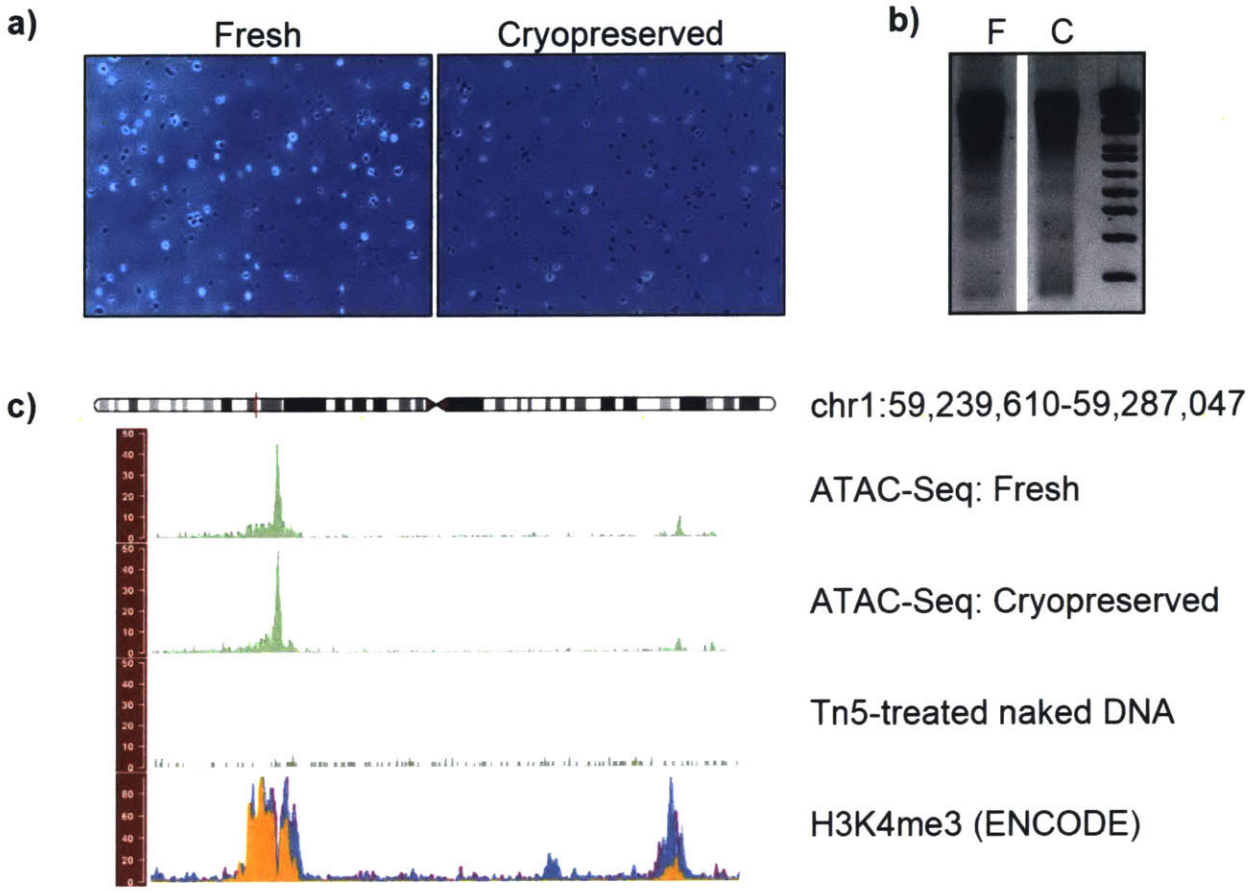

Figure 3-4 Representative results for ATAC-Seq carried out on fresh and cryopreserved cells.

(A) Nuclear morphological evaluation: similar to nuclei from fresh cells, nuclei from cryopreserved neurons were intact and of high quality. (B) Agarose gel electrophoresis of libraries: the nucleosome pattern on the gel was evident for both fresh and cryopreserved samples. (C) ATAC-Seq tracks were visualized with the Gviz package: peaks from both fresh and cryopreserved neurons were sharp and overlapped with H3K4me3 ChIP-Seq peaks from ENCODE (F = fresh, C = cryopreserved).

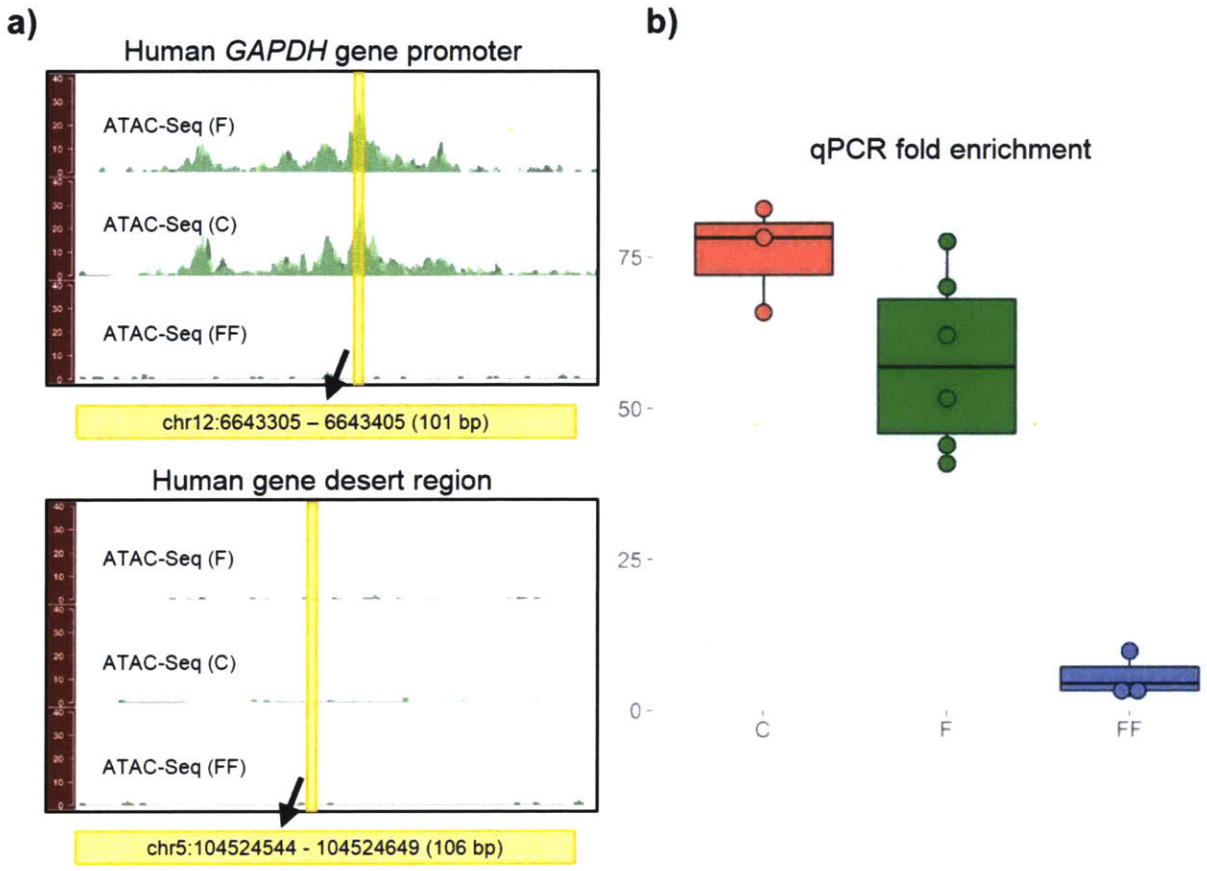

Figure 3-5 Real-time qPCR for the assessment of the quality of ATAC-Seq libraries.

(A) Genomic locations of the primers used to amplify positive (human GAPDH gene promoter) and negative (human gene desert region) control sites. (B) Fold enrichment of the open-chromatin site over the Tn5-insensitive site: while real-time qPCR experiments showed high enrichment for fresh and cryopreserved samples, poor results were obtained with flash-frozen cells. (F = fresh, FF = flash-frozen, C = cryopreserved).

Table 3-1 Information about sequencing data.

Sample	# of total reads	# of aligned reads	# of significant peaks	Fraction of reads in significant peaks (%)
F1	26,092,754	22,059,551	71,050	9.4
F2	30,730,456	25,950,925	70,073	9.1
F3	31,364,716	26,333,862	73,305	9.7
C1	28,201,642	24,577,487	73,973	10.1
C2	29,964,823	25,900,608	72,512	9.7
C3	28,018,248	23,660,777	70,547	9.2
FF	26,762,917	23,558,249	461	0.1

The numbers of total and aligned reads are indicated. The number of significant peaks is similar across fresh and cryopreserved iMNs, while only 461 peaks were detected for flash-frozen cells. The number of reads in significant peaks is > 9% for fresh and cryopreserved samples, while it is only 0.1% for flash-frozen iMNs (F = fresh, FF = flash-frozen, C = cryopreserved).

3.3.4 Quantitative comparison of fresh and cryopreserved iMNs

We subsequently performed a series of analyses to quantitatively compare the results from fresh and cryopreserved neurons. We generated sequencing data on three technical replicates from both conditions to assess whether the cryopreservation method induces any modifications in chromatin accessibility. All replicates originated from the same initial batch of cells. Information about sequencing data for each sample is reported in Table 3-1. The percentage of reads mapping to the human genome was similar for all replicates, but cryopreserved samples displayed higher number of reads mapping to mitochondrial DNA (Table 3-S2). Despite this discrepancy, we proceeded with our analysis to assess the reproducibility of the epigenetic signal from nuclear DNA across all replicates. To this purpose, we removed mtDNA reads, normalized the libraries to have the same total read counts, and examined the number of reads in 5kb genome windows (excluding ENCODE blacklisted regions). Overall, we observed high reproducibility rates ($R \geq 0.978$) between technical replicates in both fresh and cryopreserved samples (Figure 3-6A). Remarkably, cryopreserved and fresh samples were almost as highly correlated to each other ($R \geq 0.973$) as the technical replicates, which suggests that cryopreservation successfully preserves the read distribution across the genome. Next, we generated average read profiles at transcriptional start sites using the ngs.plot tool (Shen et al., 2014). As opposed to the signal from flash-frozen iMNs, highly similar patterns were observed for fresh and cryopreserved cells (Figure 3-6B). To further evaluate the similarity between cryopreserved and fresh samples, we identified the peaks in each sample and assigned each one of these peaks to neighboring features (promoters, exons, introns, distal intergenic regions and sites located downstream of the gene) within 1kb (Figure 3-6C). The distribution of peaks with respect to features in the genome was highly similar across all samples, with most of the peaks located in intergenic regions and promoters. Next, to identify and quantify potential epigenetic alterations induced by the cryopreservation procedure, we performed analysis to detect sites that were significantly different between fresh and cryopreserved samples. MACS2-derived peaks across all samples were merged into non-overlapping unique genomic intervals resulting in 75,711 sites. We then used edgeR to detect the differences between the two conditions. We identified very few

differentially enriched sites across the genome (210 out of 75,711 total = 0.28%); of these, 25.2% were located on chromosome 10, and none of them were detected on chromosome 16 (Figure 3-7A). No significant regional biases were observed for the other chromosomes. The magnitude of the differences was small, never exceeding 3-fold (Figure 3-7A and Figure 3-7B). The differentially enriched sites were mainly located in intergenic regions and promoters (57.1% and 19.5%, respectively, Figure 3-7C). We mapped 126 genes near these differentially enriched sites and performed Gene Ontology analysis using GOrilla and DAVID 6.7 tools (Eden et al., 2009; Huang et al., 2009a, 2009b). We did not detect any significant GO terms when using an adjusted p-value threshold of 0.05.

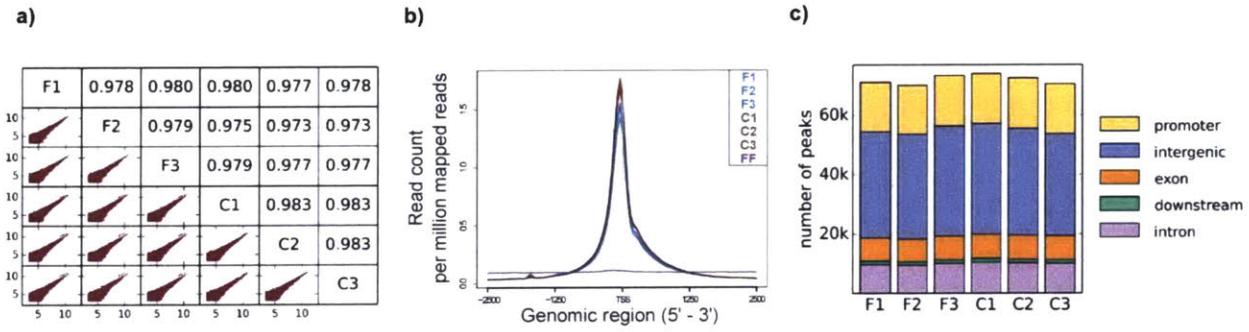

Figure 3-6 Quantitative comparison of fresh and cryopreserved cells.

(A) Correlative analysis of the number of reads in 5kb regions of the genome. The lower left triangle of the figure shows the scatter plots of the log2 read counts for each pair of technical replicates (5kb regions with less than 10 read counts were excluded from the analysis). The upper right triangle displays the corresponding values of the Pearson correlation coefficient. (B) Average read profiles across the transcriptional start sites (TSS) using a 2.5 Kb window size. The overall pattern is very similar between fresh and cryopreserved iMNs. (C) Location-based distribution analysis: the distribution of neighboring genomic features to open-chromatin regions is highly similar between fresh and cryopreserved samples. (F = fresh, FF = flash-frozen, C = cryopreserved).

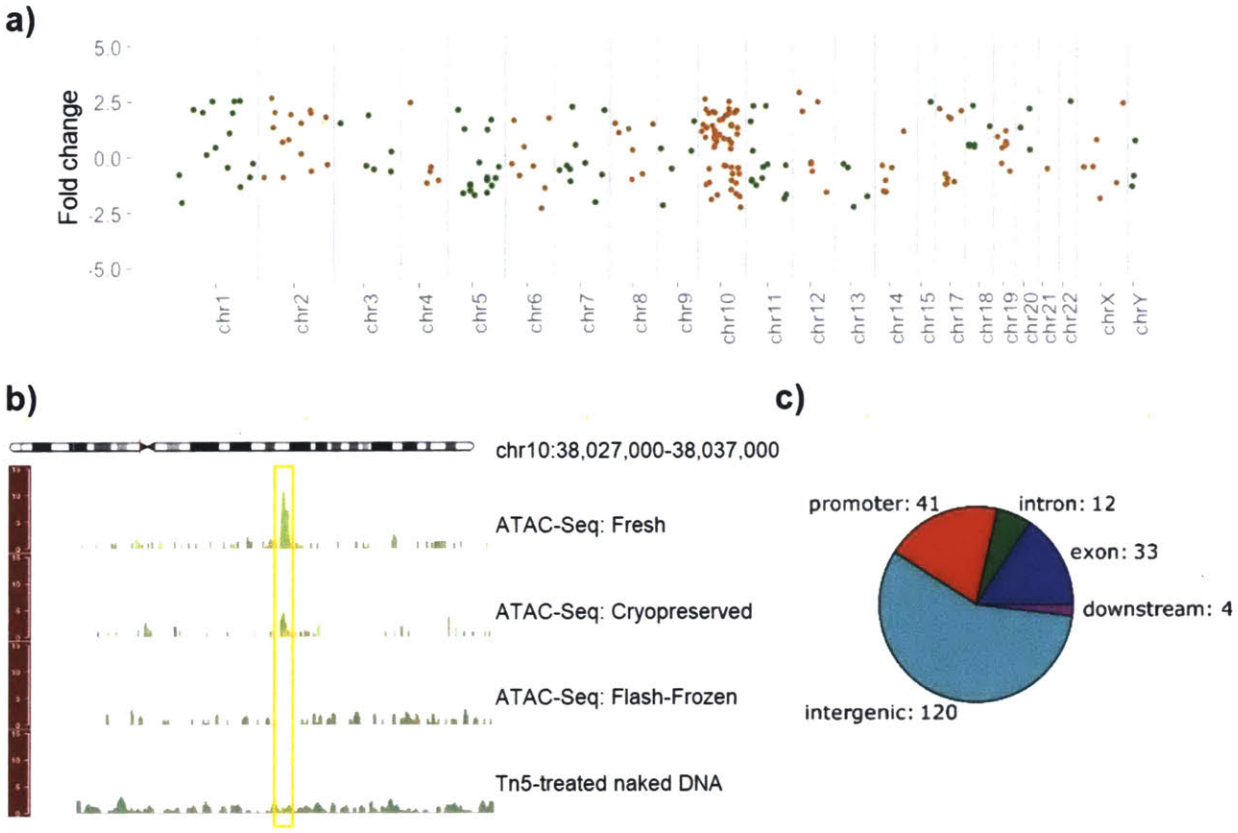

Figure 3-7 Differentially enriched sites detected between fresh and cryopreserved samples.

(A) The fold-change values for differentially enriched sites between fresh and cryopreserved samples are plotted as a function of the position of the sites across all genome. The changes were small (< 3-fold). (B) Genomic tracks of ATAC-Seq results showing a differentially enriched site between fresh and cryopreserved samples. (C) Pie chart showing the genomic location distribution of the differentially enriched sites.

In conclusion, we established a cell freezing protocol suitable for ATAC-Seq experiments on iMNs. As in the case of fresh neurons, the cryopreserved cells passed all of the quality control checkpoints. Although we observed that higher numbers of reads map to mitochondria DNA in cryopreserved iMNs, we demonstrated that the epigenetic signal from nuclear DNA was highly reproducible between fresh and cryopreserved neurons.

We expect that the method we describe also applies to a wider variety of settings and has the potential to greatly expand the number and types of samples that can be studied with ATAC-Seq. In particular, it would be interesting to test the effectiveness of this freezing procedure on additional cell types, especially heterogeneous samples such as blood-derived cells, clinical specimens, and cell co-culture systems. Indeed, different cell populations might display distinct sensitivity to freezing and thawing, with consequent biases in the epigenetic outcomes. We have described a systematic approach to assess the quality of ATAC-Seq data from frozen neurons and provided guidelines that can be followed to test the applicability of this freezing method to other sample types. We anticipate that this work will be of great value to epigenetic investigators.

3.4 Methods

Primary cells and iPSC derivation

Source fibroblast lines were obtained from Coriell (GM09677) under institutional review board approved protocols. The fibroblast-derived iPSC line 77iSMA-n5 was created by the Cedars-Sinai Medical Center iPSC Core using the episomal vectors pCXLE-hUL, pCXLE-hSK, and pCXLE-hOCT3/4-shp53-F (Addgene, from a previously published protocol (Okita et al., 2011)). We transfected the fibroblasts with the vectors using the Amaxa Human Dermal Fibroblast Nucleofector Kit. The 77iSMA-n5 line was characterized by the Cedars-Sinai iPSC Core using the following quality control assays: G-Band karyotyping, immunocytochemistry for pluripotency markers, embryoid body formation, PluriTest, and qRT-PCR for endogenous pluripotency genes (Barrett et al., 2014; Fuller et al., 2016; Müller et al., 2011; Okita et al., 2011; Sareen et al., 2012).

Motor Neuron Precursors (iMPs)

The SMA patient line, 77iSMA-n5, was grown until 90% confluent using a standard iPSC maintenance protocol. On Day 0 of differentiation, iPSCs were lifted as single cells by Accutase treatment for 5 minutes at 37ºC. We counted the cells and re-suspended them in Neuroectoderm differentiation media (NDM+LS), which contains 1:1 IMDM/F12, 1% NEAA, 2% B-27, 1% N2, 1% Antibiotic-Antimycotic, 0.2µM LDN193189 and 10µM SB431542. Next, we seeded 25,000 cells/well in a 384-well plate and centrifuged the cells for 5 minutes at 200 rcf. On day 2, we transferred the neural aggregates to a poly 2-hydroxyethyl methacrylate (poly-Hema) coated flask and cultured them for an additional 3 days in NDM+LS media. On day 5, we seeded the neural aggregates onto a tissue culture plate coated with laminin (50µg/mL) to induce rosette formation. From day 12-18, the attached neural aggregates were transitioned to Motor Neuron Specification Media (1:1 IMDM/F12, 1% NEAA, 2% B-27, 1% N2, 1% Antibiotic-Antimycotic, 0.25µM all-trans retinoic acid (ATRA), 1µM purmorphamine (PMN), 20ng/mL brain-derived neurotrophic factor (BDNF), 20ng/mL glial cell line-derived neurotrophic factor (GDNF), 200ng/mL ascorbic acid (AA) and 1µM dibutyryl cyclic-AMP (db-cAMP). On day 19 we selected the rosettes by incubating them with Neural Rosette Selection Reagent (StemCell Technologies Cat#05832) for 45 minutes at 37ºC. After selection, we collected the rosettes and transferred them to poly-Hema coated T75 flasks and cultured the cells as iMPs in Motor Neuron precursor expansion media (MNPEM), which contains 1:1 IMDM/F12, 1% NEAA, 2% B27, 1% N2, 1% Antibiotic-Antimycotic, 0.1µM ATRA, 1µM PMN, 100ng/mL EGF and 100ng/mL FGF2. We expanded the iMPs as aggregates in suspension using a mechanical passaging method known as “chopping” for up to five passages (Shelley et al., 2014; Svendsen et al., 1998). For cryopreservation, we pooled the aggregates and dissociated them via a combined enzymatic (Accutase for 10 minutes at 37°C) and mechanical dissociation strategy to form a single cell suspension. The single cell suspension was then concentrated via centrifugation (200 rcf for 5 minutes at 4°C), re-suspended in Cryostor (StemCell Technologies Cat #: 07930), cryopreserved using a controlled rate freezer (Planer Inc.) and stored in gas-phase liquid nitrogen.

Motor Neuron Cultures (iMNs)

We derived the iMNs by thawing the iMPs and immediately plating the single cell suspension onto plastic tissue culture-treated plates coated with 50µg/mL laminin for two hours at 37°C. We seeded the iMPs in Motor Neuron Maturation Medium (MNMM) Stage 1 consisting of 1:1 IMDM/F12, 1% NEAA, 2% B-27, 1% N2, 1% Antibiotic-Antimycotic, 0.1µM ATRA, 1µM PMN, 10ng/mL BDNF, 10ng/mL GDNF, 200ng/mL AA, 1µM db-cAMP, and 2.5µM N-[(3,5-Difluorophenyl)acetyl]-L-alanyl-2-phenyl]glycine-1,1-dimethylethyl ester (DAPT). We cultured the cells for a period of seven days. On day 7, the plated cultures were transitioned to MNMM Stage 2 containing 98.8% Neurobasal media, 1% non-essential amino acids, 0.5% Glutamax, 1% N2, 10ng/mL BDNF, 10ng/mL GDNF, 200ng/mL AA, 1µM db-cAMP, and 0.1µM Ara-C. We further differentiated the iMNs in MNMM Stage 2 for a total of 21 days. On day 21, the iMNs cultures were either fixed for immunocytochemistry or collected.

Cell collection, freezing, and thawing

For cell collection, the iMNs were washed once with 1X PBS, isolated via cell scraper in 1X PBS, and centrifuged at 200 rcf for 5 minutes at 4°C. Aliquots with approximately one million cells were prepared for each experimental condition. Flash-freezing: pellets (no supernatant) were flash-frozen in liquid nitrogen. Cryopreservation: pellets were re-suspended in Cryostor media and frozen slowly in a Mr. Frosty isopropyl alcohol chamber (FisherSci) at -80°C. This procedure allowed us to achieve a rate of cooling of -1°C/minute.

Both the flash-frozen isolated cell pellets and the cryopreserved iMNs were stored for 10 days at -80°C. To thaw the cryopreserved iMNs, we removed the cryovials from -80°C and quickly warmed them for 2 minutes in a 37°C water bath. We transferred the samples to 12ml of warm 1X PBS supplemented with 1X protease inhibitor cocktail. We gently mixed each tube by inversion and removed an aliquot (100µl) for cell death estimation using the chromatin condensation assay described below. We centrifuged the cells at 200 rcf for 5 minutes at 4°C, carefully aspirated all the supernatant and proceeded with nuclei isolation. Flash-frozen cell pellets were removed from -80°C and immediately re-suspended in ice-cold cell lysis buffer.

Chromatin condensation assay

Hoechst 33342 was added to 100µl of cell suspension at a final concentration of 1.5µg/ml. Cells were incubated for 15 minutes at 37°C before proceeding with image acquisition which was carried out using a 350nm UV excitation filter. Eight randomly selected fields per sample were imaged at 40X magnification. Neurons were scored as apoptotic when they showed condensed chromatin or fragmented nuclei with bright Hoechst signal.

Immunocytochemistry

We fixed iMNs with 4% paraformaldehyde and blocked them with 5% donkey serum with 0.1% Triton X-100 in 1X PBS. We incubated the cells overnight at 4°C with the following primary antibodies: anti-SMI32 (mouse monoclonal, 1:1,000, BioLegend, cat. no. SMI-32R) and anti-ISL1 (goat polyclonal, 1:250, R&D Systems, cat. no. AF1837). We subsequently rinsed the cells and incubated them with species-specific Alexa Fluor 488-conjugated secondary antibody (donkey anti-mouse immunoglobulin G (IgG), 1:1,000, Life Technologies, cat. no. A-21202) and Alexa Fluor 594-conjugated secondary antibody (donkey anti-goat IgG, 1:1000, Life Technologies, cat. no. A-11058). We counterstained nuclei using DAPI (1µg/mL). We acquired the images using Nikon/Leica microscopes with 10x magnification.

Purification of nuclei from iMNs

We re-suspended the cell pellets in ice-cold cell lysis buffer (10mM Tris-HCl, pH7.4, 10mM NaCl, 3mM MgCl2, 0.1% IGEPAL CA-630) supplemented with 1X protease inhibitor cocktail (Roche). We incubated the cells on ice for 5 minutes and centrifuged at 230 rcf for 5 minutes at 4°C. We carefully removed the supernatant and re-suspended the nuclei in 25µl of ice-cold 1X Tagment DNA Buffer (Illumina). We quantified the nuclei with Trypan Blue staining and the Countess® Automated Cell Counter (Invitrogen).

DNA extraction

We purified the DNA from iMNs using the DNeasy Blood & Tissue Kit (Qiagen), according to the manufacturer's instructions. We quantified the DNA using a NanoDrop 2000 instrument (Thermo Scientific) and used 50ng to prepare the DNA library using the Nextera DNA Library Preparation Kit (Illumina), according to the manufacturer's instructions. This library, obtained from naked DNA, was used as internal control to determine the background level of intrinsic accessibility of genomic DNA and correct for any Tn5 transposase sequence cleavage bias.

Chromatin tagmentation and sequencing

We used 50,000 nuclei for the transposase reaction, which was carried out as described in Buenrostro et al. (Buenrostro et al., 2013). We subsequently purified the samples with the DNA Clean & Concentrator–5 Kit (Zymo Research) and eluted them with 20µl of Elution Buffer (Qiagen). We PCR-amplified the samples using 25µl of Nextera PCR Master Mix (Illumina), 5µl of PCR Primer Cocktail (Illumina), 5µl of Index primer 1 (i7, Illumina), and 5µl of Index primer 2 (i5, Illumina). We used the following PCR reaction protocol: 3min 72°C; 30sec 98°C; 8 cycles (10sec 98°C, 30sec 63°C, 3min 72°C). We purified the samples with the DNA Clean & Concentrator–5 Kit (Zymo Research), eluted them with 20µl of Elution Buffer (Qiagen), and loaded them on 2% agarose gel (Invitrogen) for qualitative evaluation of libraries and size-selection. We size-selected the following fractions: 175 - 250 bp (fraction "A", corresponding to a nucleosome-free fragment size) and 250 - 625 bp (fraction "B"). We purified the DNA from both gel fractions, using the QIAquick Gel Extraction Kit (Qiagen) following the manufacturer's recommendation, and eluted it with 20µl of Elution Buffer (Qiagen). We utilized the DNA from fraction "B" for qPCR-based qualitative analysis of libraries using primers mapping to open-chromatin regions as positive control sites and gene desert regions as negative control sites (Figures 3-5 and 3-S4). The sequences of the primers used to amplify open-chromatin and gene desert regions are shown in Table 3-S3. We also performed the qPCR assay using 10-fold serial dilutions of non-transposed genomic DNA as a template to generate a calibration line for each primer pair and correct for any differences in the primer efficiency. The fold enrichment of the open-

chromatin site (OC) over the Tn5-insensitive site (INS) was calculated with the following formula, as previously described: 2 to the power of [(OCn-OCa) - (INSn-INSa)], where OCn is the qPCR threshold cycle number obtained for the OC qPCR primer pair using transposed naked DNA as template, and INSa is the qPCR threshold cycle number obtained for the INS qPCR primer pair using ATAC-Seq library as template (Ling and Waxman, 2013). As an additional control, we carried out the qPCR assay using transposed naked DNA. No fold-enrichment of open-chromatin sites should be detected when using transposed naked DNA as a template. We prepared the amplification reaction with 1X KAPA SYBR FAST qPCR Master Mix (Kapa Biosystems) and 500nM of forward and reverse primers. We carried out qPCR assays using a LightCycler® 480 Instrument II (Roche), available at the MIT BioMicroCenter. We further amplified the DNA from fraction "A" with 1X NEBNext High-Fidelity PCR Master Mix (New England Biolabs), 200nM of Primer 1 (5'-AATGATACGGCGACCACCGA-3'), and 200nM of Primer 2 (5'-CAAGCAGAAGACGGCATACGA-3'). We used the following PCR reaction protocol: 30sec 98°C; 4 cycles (10sec 98°C, 30sec 65°C, 30sec 72°C); 5min 72°C. We purified the final libraries using Agencourt AMPure XP beads (Beckman Coulter), checked their quality using a Fragment Analyzer™ instrument (Advanced Analytical), and measured their concentration by a qPCR-based method (KAPA Library Quantification Kit for Illumina Sequencing Platforms). We submitted the samples to the MIT BioMicroCenter for single-end sequencing with the Illumina HiSeq 2000 platform.

Bioinformatic analysis

We aligned sequencing reads to the hg19 genome build using BWA v.0.7.10. We assessed the quality of the sequences using FastQC (more details on how the data was processed can be found at http://openwetware.org/wiki/BioMicroCenter:Software#BMC-BCC_Pipeline). Given the large percentage of mitochondrial reads found in some samples, we removed mitochondrial reads from the analysis using custom UNIX scripts. We determined open-chromatin regions (peaks) using MACS2 v.2.1.0.20150420 (q-value threshold = 0.05) (Zhang et al., 2008). We used the sequencing data from transposed naked DNA as a control. The differential analysis was performed using the default settings in the package DiffBind version 1.16.0 (Ross-Innes et al., 2012). Briefly,

read counts for each site were computed and differentially enriched sites between fresh and cryopreserved conditions were identified using the edgeR package, with FDR < 0.1 (Robinson et al., 2009).

3.5 Supplementary Information

Table 3-S1 Information about the number of cells used for the experiment, the percentage of cell death assessed by chromatin condensation and the number of nuclei recovered from cryopreserved (C) neurons.

Sample	# of cells	Cell death (%)	# of recovered nuclei
C1	947,150	11.0	762,000
C2	1,282,500	8.3	921,000
C3	1,225,500	12.3	849,000

Table 3-S2 Mitochondrial DNA (mtDNA) contamination in fresh (F) and cryopreserved (C) iMNs.

Sample	mtDNA (%)
F1	32.14
F2	27.97
F3	31.44
C1	45.33
C2	49.61
C3	47.69

Table 3-S3 Sequences of the primers used to amplify open-chromatin and gene desert regions.

Primer ID	Primer sequence
GAPDH gene promoter Fw	CATCTCAGTCGTTCCCAAAGT
GAPDH gene promoter Rv	TTCCCAGGACTGGACTGT
Gene desert region Fw	AACTGGCTAGTAAGGAGTGAATG
Gene desert region Rv	GGGAATGGAAAGAAGTCCACTAT
B2M gene promoter Fw	GGAAAGTCCCTCTCTCTAACCT
B2M gene promoter Rv	GCGACGCCTCCACTTATATT
Gene desert region #2 Fw	CCCAAACTCTGAGAGGCTTATT
Gene desert region #2 Rv	GAGCCATCATCTAGACACCTTC

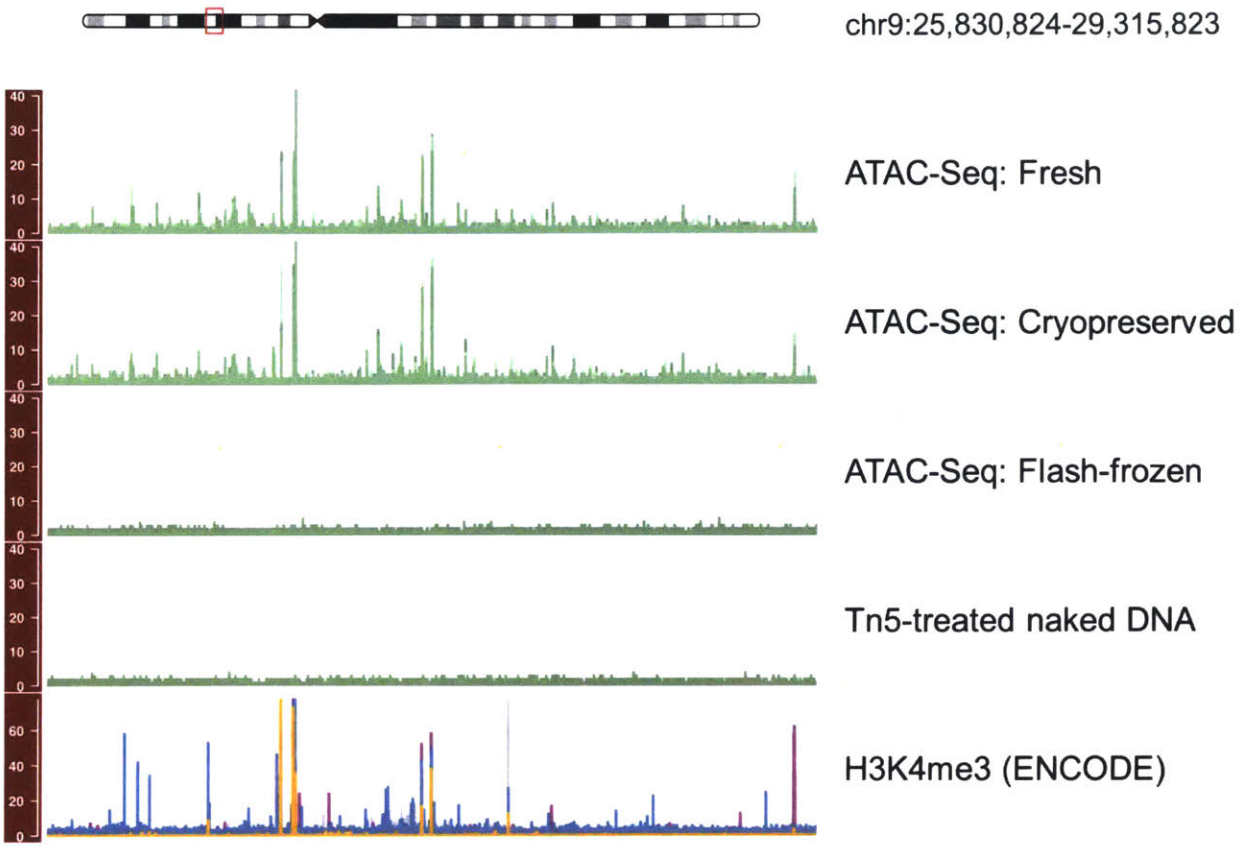

Figure 3-S1 ATAC-Seq tracks of a large genomic region (3.5 Mbp).

The tracks were visualized with the *Gviz* package: peaks from both fresh and cryopreserved neurons were sharp and overlapped with H3K4me3 ChIP-Seq peaks from ENCODE; the reads from flash-frozen neurons were distributed noisily across the genome (F = fresh, FF = flash-frozen, C = cryopreserved).

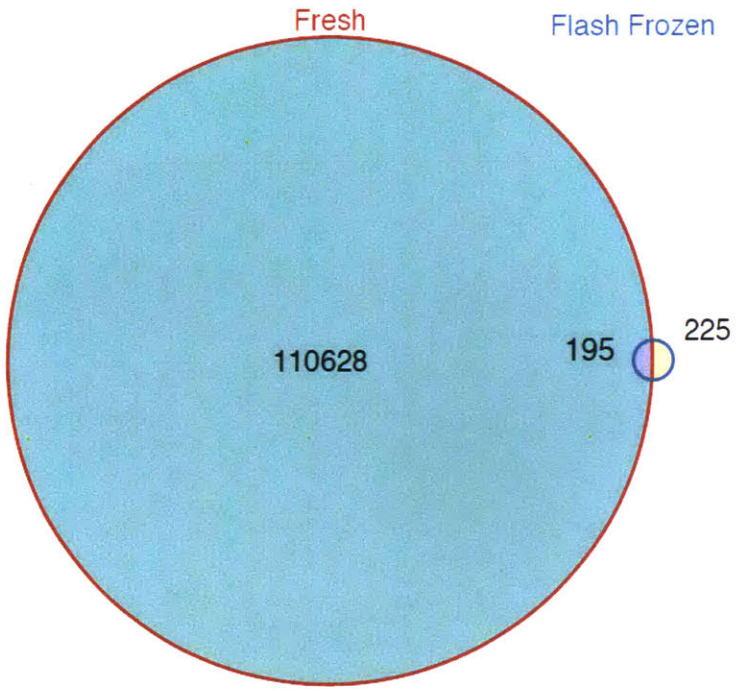

Figure 3-S2 Venn diagram showing the overlap of the peaks between fresh and flash-frozen iMNs.

The reads from the three technical replicates from the fresh iMNs were merged before calling the peaks with MACS2 and calculating the overlap with the peaks from flash-frozen iMNs. We observed that 236 out of 461 peaks detected in the flash-frozen iMNs overlapped with the peaks obtained from the fresh cells. In some cases, multiple peaks from a sample mapped to a single peak in the second sample. Such ties were counted as a single overlap, resulting in the 195 overlapping peaks displayed on the Venn diagram.

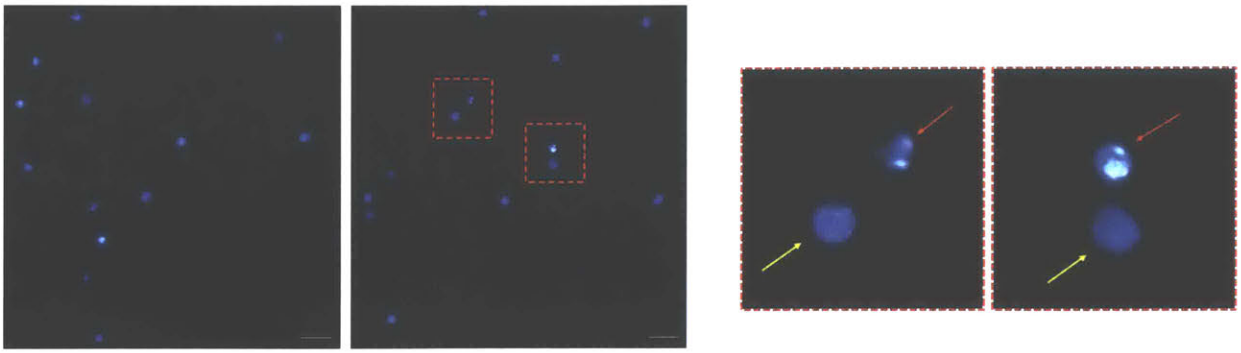

Figure 3-S3 Two representative microscopic pictures of thawed cells stained with Hoechst 33342 for the assessment of neuronal death based on chromatin condensation.

The right panel shows the corresponding enlarged images from the left panel. The red arrows indicate apoptotic cells with condensed and fragmented chromatin and bright Hoechst signal, while the yellow arrows indicate viable cells with diffuse staining. Scale bar = 40 μm.

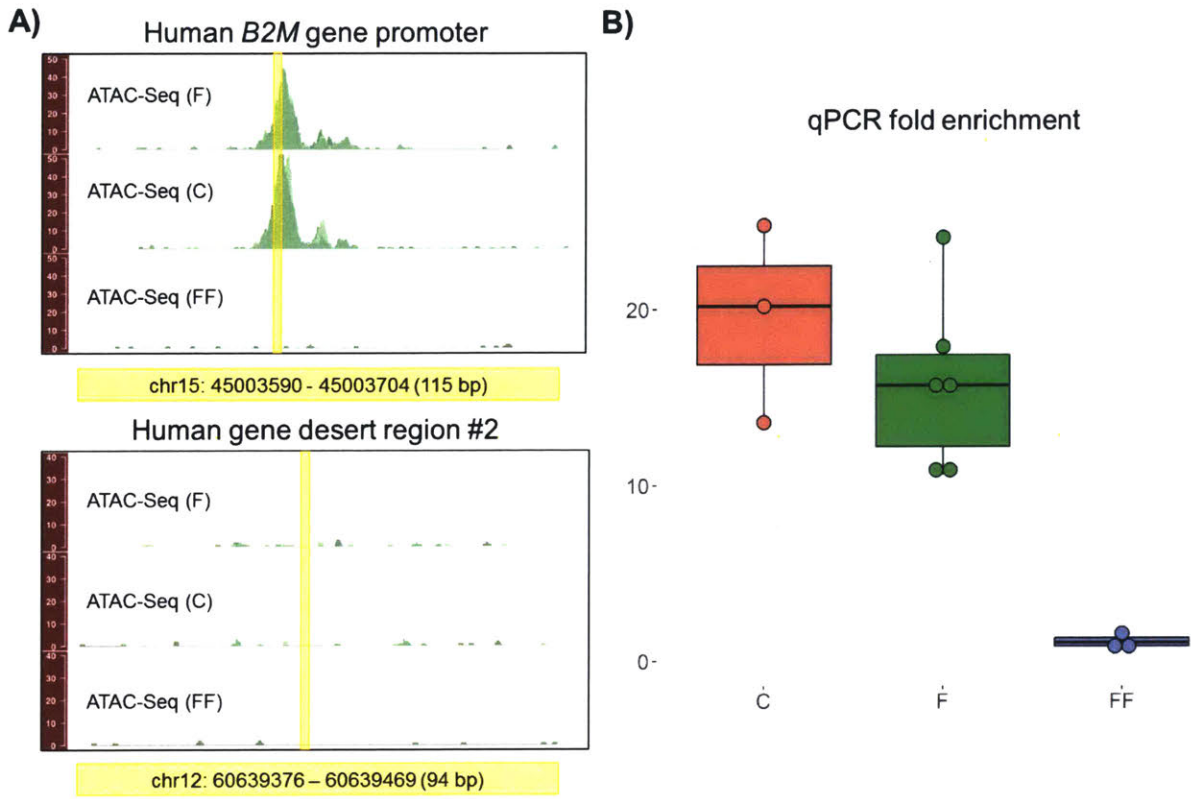

Figure 3-S4 Real-time qPCR for the assessment of the quality of ATAC-Seq libraries.

(A) Genomic locations of the primers used to amplify positive (human *B2M* gene promoter) and negative (human gene desert region) control sites. (B) Fold enrichment of the open-chromatin site over the Tn5-insensitive site: while real-time qPCR experiments showed high enrichment for fresh and cryopreserved samples, poor results were obtained with flash-frozen cells (F = fresh, FF = flash-frozen, C = cryopreserved).

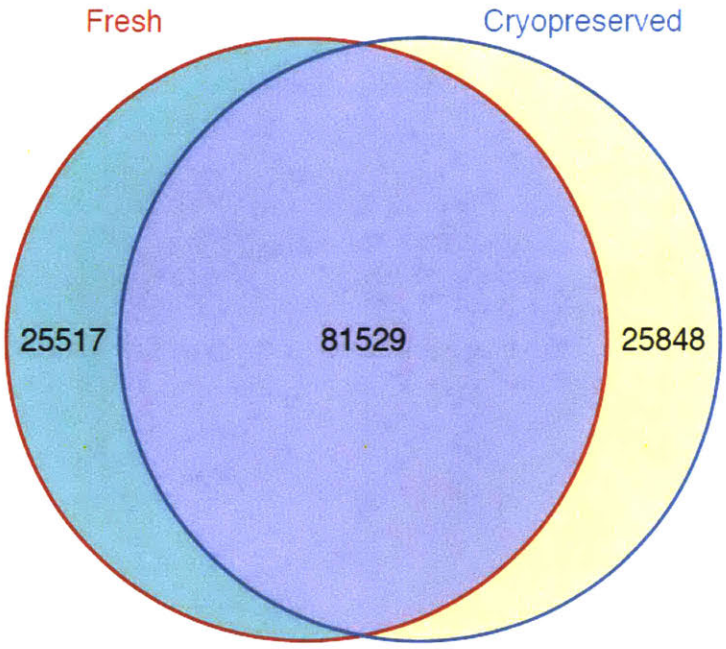

Figure 3-S5 Venn diagram showing the overlap of the peaks between fresh and cryopreserved iMNs.

The reads from the three technical replicates from both fresh and cryopreserved iMNs were merged before calling the peaks with MACS2 and calculating the overlap between the two conditions.

3.6 References

Barrett, R., Ornelas, L., Yeager, N., Mandefro, B., Sahabian, A., Lenaeus, L., Targan, S.R., Svendsen, C.N., and Sareen, D. (2014). Reliable Generation of Induced Pluripotent Stem Cells From Human Lymphoblastoid Cell Lines. Stem Cells Transl. Med. *3*, 1429–1434.

Brind'Amour, J., Liu, S., Hudson, M., Chen, C., Karimi, M.M., and Lorincz, M.C. (2015). An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat. Commun. *6*, 6033.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015a). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015b). ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol. *109*, 21–29.

Dovey, S.L. (2012). Oocyte cryopreservation: Advances and drawbacks. Minerva Ginecol. *64*, 485–500.
Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.
Fuller, H.R., Mandefro, B., Shirran, S.L., Gross, A.R., Kaus, A.S., Botting, C.H., Morris, G.E., and Sareen, D. (2016). Spinal Muscular Atrophy Patient iPSC-Derived Motor Neurons Have Reduced Expression of Proteins Important in Neuronal Development. Front. Cell. Neurosci. *9*.
Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.
Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.
Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T.M., Childs, R., et al. (2015). Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature *528*, 142–146.
Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Chromatin state dynamics during blood formation. Science *345*, 943–949.
Ling, G., and Waxman, D.J. (2013). DNase I digestion of isolated nulcei for genome-wide mapping of DNase hypersensitivity sites in chromatin. Methods Mol. Biol. *977*, 21–33.
Müller, F.J., Schuldt, B.M., Williams, R., Mason, D., Altun, G., Papapetrou, E.P., Danner, S., Goldmann, J.E., Herbst, A., Schmidt, N.O., et al. (2011). A bioinformatic assay for pluripotency in human cells. Nat. Methods *8*, 315–317.
Ogino, S., and Wilson, R.B. (2004). Spinal muscular atrophy: Molecular genetics and diagnostics. Expert Rev. Mol. Diagn. *4*, 15–29.
Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K.I., et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat. Methods *8*, 409–412.
Paramanantham, J., Talmor, A.J., Osianlis, T., and Weston, G.C. (2015). Cryopreserved oocytes: Update on clinical applications and success rates. Obstet. Gynecol. Surv. *70*, 97–114.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.
Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389–393.
Sareen, D., Ebert, A.D., Heins, B.M., McGivern, J. V., Ornelas, L., and Svendsen, C.N. (2012). Inhibition of apoptosis blocks human motor neuron cell death in a stem cell model of spinal muscular atrophy. PLoS One *7*, e39113.
Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G., and Greenleaf,

W.J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. *25*, 1757–1770.
Schmidl, C., Rendeiro, A.F., Sheffield, N.C., and Bock, C. (2015). ChIPmentation: Fast, robust, low-input ChIP-seq for histones and transcription factors. Nat. Methods *12*, 963–965.
Shelley, B.C., Gowing, G., and Svendsen, C.N. (2014). A cGMP-applicable Expansion Method for Aggregates of Human Neural Stem and Progenitor Cells Derived From Pluripotent Stem Cells or Fetal Brain Tissue. J. Vis. Exp.
Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). Ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.
Svendsen, C.N., Ter Borg, M.G., Armstrong, R.J.E., Rosser, A.E., Chandran, S., Ostenfeld, T., and Caldwell, M.A. (1998). A new method for the rapid and long term growth of human neural precursor cells. J. Neurosci. Methods *85*, 141–152.
Trusal, L.R., Guzman, A.W., and Baker, C.J. (1984). Characterization of freeze-thaw induced ultrastructural damage to endothelial cells in vitro. In Vitro *20*, 353–364.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.
Ziegler, U., and Groscurth, P. (2004). Morphological Features of Cell Death. Physiology *19*, 124–128.

Chapter 4: Conclusion

4.1 Summary and implications

This thesis describes a multi-omics approach to understanding the molecular effects of perturbagens in models of neurodegenerative disorders. Elucidating the systemic changes induced by compounds or genetic manipulations is a major challenge in disease research, both in basic science and drug discovery. In Chapter 2, we explored chemical perturbagens in a Huntington's Disease (HD) model and identified novel modes of action (MoAs). In Appendix A, we discovered the effects of two methods of huntingtin silencing in mouse liver tissue. In Chapter 3 and Appendix B, we turned our attention to motor neuron diseases, particularly Spinal Muscular Atrophy (SMA) and Amyotrophic Lateral Sclerosis (ALS), and identified the cellular processes affected by particular genetic states. The projects presented in this thesis have implications for academia and industry.

Chapter 2 represents the bulk of my graduate work. We developed a multi-omics, machine learning approach for identifying the MoAs of chemical perturbagens without the need for reference compounds or specific knowledge about regulatory interactions. To demonstrate the utility of this approach, we sought to identify MoAs for compounds identified in the search for drugs to treat HD, an invariably fatal neurodegenerative disorder. More than a hundred such compounds have been identified, but so far, none have succeeded to modify disease progression in clinical trials (Kumar et al., 2015; Zuccato et al., 2010).

We gathered multi-omics data, including RNA-Seq, metabolomics, H3K4me3 ChIP-Seq and proteomics, from HD cells treated with a subset of these compounds. Surprisingly, we show that previously unrelated compounds cluster together in some omics data. Importantly, these unexpected groupings may occur using one type of omics data, but not another. In the particular cases we examined, these groupings suggest shared MoAs that would not be expected based on similarities in the compounds' screening results, structures, Connectivity Map connectivity scores, or known binding targets alone. For instance, two of the compounds are known antagonists of the histamine H1 receptor, yet they belong to different clustering groups in our metabolomic and proteomic data (Wishart et al., 2018a). To find the underlying

MoAs for each group, we used a feature selection approach that leverages known molecular interactions from public databases (Brunk et al., 2018; Hornbeck et al., 2015; Pirhaji et al., 2016; Razick et al., 2008; Wishart et al., 2018b). A machine-learning network optimization algorithm applied to this interactome reveals the altered cellular processes (Tuncbag et al., 2016).

Crucially, we experimentally validated the most HD-relevant MoAs. We show, for example, that one of the two antihistamines that ameliorates HD phenotypes has a profound effect on autophagy. Based on the literature and its transcriptional profile, the other antihistamine would have been assumed to have the same MoA. Surprisingly, our data show that this second antihistamine actually has no effect on autophagy. Instead, it targets bioenergetics. In this example, two compounds with potential benefit for the same disease and with the same reported target actually functioned through completely distinct MoAs. We also found an example where two compounds had the same MoA despite having little similarity in chemical structure and no common binding target. Specifically, we show that an inhibitor of diacylglycerol kinase (DKI) had previously unknown effects on mitochondrial respiration, ATP production, and glycolysis in a similar manner to the antihistamine meclizine. These novel MoAs can be used in the context of other diseases where the specified effect is needed. The general machine learning approach can also be used for compounds in other systems to identify their MoAs. The results from this project can impact future drug repurposing and drug development efforts.

We also profiled the effects of genetic perturbagens in the context of HD in Appendix A. In collaboration with Dr. Jeff Carroll, we compared the effects of two huntingtin gene silencing methods on the transcriptome and metabolome of mouse liver. One silencing method involved the use of antisense oligonucleotides (ASOs) and the other used gene knockouts with cell-type specificity for hepatocytes (Coffey et al., 2017). We found that both silencing techniques had a significant influence on gene expression, but little impact on metabolite abundance. They also affected similar cellular pathways, such as immune response and fatty acid metabolism. It is imperative to understand the peripheral effects of huntingtin silencing because there are ongoing clinical trials that use ASOs to silence huntingtin in humans with HD.

Genetic perturbagens were also explored in the contexts of SMA and ALS in collaboration with the NeuroLINCS consortium. In Chapter 3, we developed a cell freezing protocol suitable for ATAC-Seq on motor neurons derived from induced pluripotent stem cells derived from patients with SMA. These cells have a deficiency in the *SMN1* gene. We found that cryopreserved cells retained their chromatic structure, unlike flash-frozen cells, and could be used for ATAC-Seq. This work informed the community about the procedure to keep chromatin intact in precious samples for future studies. In Appendix B, we characterized similar motor neurons, but these were derived from patients with ALS and carry hexanucleotide expansions in *C9orf72*. Multi-omics network analysis was performed to identify the changes induced by this genetic perturbagen. Fly screen data from an ALS fly model was used to label affected cellular pathways as causal or compensatory in ALS. These labels can inform future drug targeting efforts.

4.2 Limitations and future perspectives

In this thesis, we used systems biology approaches to examine the multi-omics effects of perturbagens. These approaches can be extended to understand the effects of drugs or diseases in other contexts. However, it is important to note that the field is constantly evolving and there are limitations in these studies.

One limitation is the use of model systems that do not fully capture the complex pathophysiology of neurodegenerative disorders. The purpose of applying chemical and genetic perturbagens to these models is to understand the disease response or the behavior of the perturbagen itself, with the hopes that the responses will mimic those in humans. In Chapter 2, we used the murine STHdh^{Q111} cell line model of HD, as well as human SH-SY5Y and HEK293 cell lines. Though we were able to identify modes of action, the results from immortalized cell lines might not recapitulate the effects the compounds could have in humans (Trettel et al., 2000). Unknown or unmodeled interactions between cell types and tissues could affect the response of the compounds. In Appendix A, a mouse model was used to understand the complex effects of peripheral huntingtin silencing. In this model, only the liver tissue was profiled. The translatability of the results could be hampered due to the differences between humans

and mice. Also, compensatory interactions could confound the results, as the liver interacts with other organs within the mice and these effects were not profiled. In Chapter 3 and Appendix B, motor neurons differentiated from induced pluripotent stem cells from human SMA and ALS patients, respectively, were profiled. Though these cells are derived from human patients, the homogeneity of the differentiated cultures could confound the results. The dependencies between motor neurons and other cells in the brain could lead to unexpected effects (Sances et al., 2016).

With developments and advances in biological data collection, it is likely that the field will turn to single-cell omics data to understand the specific responses in distinct cell types. Models of neurodegenerative disorders are also constantly being improved. Protocols for developing isogenic neuronal cultures derived from humans will be important for preclinical research. These cultures should mimic the neuronal cell death phenotype present in the human diseases and could be used as the gold standard model for disorders such as HD, SMA, and familial ALS.

Though all models of neurodegenerative disorders will introduce confounding effects and will not comprehensively mimic actual human disease, they are necessary to study in preclinical research before testing any perturbagens in actual humans. By studying the effects of chemical perturbagens in multiple models, like we did for autophagy in Chapter 2, we can better understand the consistency of MoAs. Future work can also use the identified MoAs in other disease contexts for the purpose of drug repurposing. If a compound has the same effect in multiple disease models spanning different organisms, we can perhaps expect it to work in a similar manner in humans. Along with testing the perturbagens in multiple models and examining additional compounds, multiple treatments times and doses for compounds should be considered to understand the systemic responses to the compounds.

As technology improves and the field progresses, omics assays will become cheaper and more routine. Systems biology has already seen a dramatic increase in omics data generation. Beyond the reduced-representation data in the Connectivity Map and LINCS databases, libraries of detailed omics effects of compounds will likely be developed, and these libraries could be used both in academia and industry initiatives. Like compound libraries that are available for chemical testing, omics libraries could be

created and downloaded as a standardized package for analysis. New tools that can combine data across labs or batches will be important to standardize the data collected. The field will turn to general-use methods that integrate different types of data in an unbiased manner to determine the functional effects of compounds. The results of these efforts would bring together more collaborations between basic scientists and industry researchers. The increased knowledge of pathways and how they can be modulated would be of interest in basic science, and the identification of compounds with similar modes of action would be of interest in industry. Overall, the omics libraries could lead to more efficient drug repurposing and fewer side effect surprises.

Another limitation to the studies in this thesis involves the systems biology approaches we have administered. The multi-omics integrative network approaches depended on high-quality omics data and a dependable interactome. As more interactions or specificities of interactions are discovered, better interactomes can be created to answer different biological questions. For example, interactions that occur in specific cell types or tissues could be excluded when those cell types or tissues are not under consideration. Also, as more information is learned about protein-metabolite interactions, better confidence scores and edge costs can be given to those interactions, which would improve the identification of functional pathways.

Once pathways were identified to be affected by the perturbagens, we had to experimentally validate the changes. The machine learning methods used in this thesis do not directly predict causal changes or the precise changes in pathways that lead to the overall functional changes without further experimental testing. Future work to incorporate causal modeling and prediction in our integrative multi-omics methods is needed to reduce the amount of experimental testing required. Some methods, including Bayesian network modeling and structural equations modeling, have been proposed to predict causal relationships between molecules in a network (Auerbach et al., 2018; Sachs, 2005). This would improve our understanding of the directionality of the multiple changes induced by a perturbagen in a cellular pathway. Also, these predictive methods could be used to examine the molecules within a pathway and rank targets that would otherwise be overwhelming and confusing to choose from when embarking on experiments.

Advances in gene silencing has spurred the emergence of HD clinical trials focused on silencing huntingtin using ASOs. The field has high hopes for these clinical trials because they directly target the cause of HD instead of just treating downstream symptoms. The two early-stage trials currently in progress include allele-specific and non-allele-specific targeting of huntingtin. The allele-specific ASO targets single nucleotide polymorphisms in the mutated huntingtin allele, while the non-allele-specific ASO targets both the mutated and wild-type huntingtin alleles. The current approach for ASO delivery to the brain involves intrathecal administration, which is a painful and invasive technique. Future work studying the effects of huntingtin ASOs in humans will be necessary to determine possible unwanted side effects. Omics profiling of cells derived from humans treated with huntingtin ASOs will also be helpful to understand the specific pathways that would need to be modulated to achieve a similar overall effect. If the clinical trials succeed, the field will likely head in the direction of improving ASO drug delivery or finding alternative drugs. By comparing the omics profiles of the silenced cells to those treated with various chemical perturbagens, perhaps small molecules can be found that affect the same pathways in the desired manner.

Because neurodegenerative disorders are so complex, a combination of multiple perturbagens will likely be necessary to achieve an effective disease-modifying response. The field will probably turn to synergy modeling and testing to understand how multiple small molecules could be combined to specify and fine-tune responses in multiple cellular pathways. Many small molecules have been proposed to treat HD, but they have not had much success. Based on the relative success in the field of cancer research, we may begin to see cocktails of various drugs in clinical trials for neurodegenerative disorders.

Overall, this thesis has illustrated the power of using systems biology approaches to understand the effects of perturbagens in neurodegenerative disease research. We studied chemical perturbagens to identify MoAs that can be used in drug discovery efforts. We also examined genetic perturbagens can be used to understand disease mechanisms. The MoAs and cellular disease processes identified can be used to guide future therapies.

4.3 References

Auerbach, J., Howey, R., Jiang, L., Justice, A., Li, L., Oualkacha, K., Sayols-Baixeras, S., and Aslibekyan, S.W. (2018). Causal modeling in a multi-omic setting: insights from GAW20. BMC Genet. *19*, 74.

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. *36*, 272–281.

Coffey, S.R., Bragg, R.M., Minnig, S., Ament, S.A., Cantle, J.P., Glickenhaus, A., Shelnut, D., Carrillo, J.M., Shuttleworth, D.D., Rodier, J.-A., et al. (2017). Peripheral huntingtin silencing does not ameliorate central signs of disease in the B6.HttQ111/+ mouse model of Huntington's disease. PLoS One *12*, e0175968.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512-20.

Kumar, A., Kumar Singh, S., Kumar, V., Kumar, D., Agarwal, S., and Rana, M.K. (2015). Huntington's disease: An update of therapeutic strategies. Gene *556*, 91–97.

Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C.B., White, F.M., Saghatelian, A., and Fraenkel, E. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. Nat. Methods *13*, 770–776.

Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.

Sachs, K. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science (80-.). *308*, 523–529.

Sances, S., Bruijn, L.I., Chandran, S., Eggan, K., Ho, R., Klim, J.R., Livesey, M.R., Lowry, E., Macklis, J.D., Rushton, D., et al. (2016). Modeling ALS with motor neurons derived from human induced pluripotent stem cells. Nat. Neurosci. *19*, 542–553.

Trettel, F., Rigamonti, D., Hilditch-Maguire, P., Wheeler, V.C., Sharp, a H., Persichetti, F., Cattaneo, E., and MacDonald, M.E. (2000). Dominant phenotypes produced by the HD mutation in STHdh(Q111) striatal cells. Hum. Mol. Genet. *9*, 2799–2809.

Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput. Biol. *12*, e1004879.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018a). DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074-82.

Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., et al. (2018b). HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Res. *46*, D1074-82.

Zuccato, C., Valenza, M., and Cattaneo, E. (2010). Molecular Mechanisms and Potential Therapeutical Targets in Huntington ' s Disease. Physiol Rev *90*, 905–981.

Appendix A: Molecular Effects of Huntingtin Silencing in Mouse Liver

This work is being prepared for publication.

This project is part of a collaboration with Professor Jeff Carroll at Western Washington University, whose lab developed the mice models and collected the transcriptomic and metabolomic data.

As part of this work, I downloaded DNAse-Seq data from ENCODE and would like to acknowledge the ENCODE Consortium and the ENCODE production laboratory of Professor John Stamatoyannopoulos.

My contributions:
I analyzed the transcriptomic and metabolomic data for the different mouse models of huntingtin silencing.

A.1 Introduction

Huntington's Disease (HD) is a fatal neurodegenerative disease caused by abnormal expansion of a CAG repeat in the huntingtin gene (Tabrizi et al., 2019). The resultant mutated protein is ubiquitously expressed and causes several loss of function and toxic gain of function mechanisms (Zuccato et al., 2010). There is no cure for the disease, but recent clinical development has focused on huntingtin lowering strategies to reduce the pathogenic effects of the mutated protein.

There are currently two ongoing clinical trials for huntingtin-lowering therapies, and both use RNA-targeting antisense oligonucleotide (ASO) approaches. The ASOs are delivered intrathecally and have different allele selection criteria. The first trial, sponsored by Ionic Pharmaceuticals, is currently in phase 3 enrollment and features huntingtin gene silencing without allele specificity, targeting both the wild-type and mutant alleles (Tabrizi et al., 2019). The second trial, sponsored by Wave Life Sciences, is currently in phase 1b/2a and features selective silencing of only the mutant allele by using SNP targets around the mutated repeat region of the gene (Tabrizi et al., 2019). As the trials are still in the early stages, many questions regarding huntingtin lowering remain unanswered.

Preclinical research in animal models is critical to better understand the effects of huntingtin silencing. Most studies have focused on the tolerability of various huntingtin lowering strategies in the brain (Kaemmerer and Grondin, 2019). However, few have identified the effects of huntingtin silencing in peripheral tissues. ASOs delivered to the brain may leak into peripheral circulation and it has been hypothesized that intact ASOs could accumulate in peripheral organs (Jeff Carroll, unpublished work). After intrathecal delivery of ASOs in human studies, silencing of huntingtin could occur in peripheral tissues. Knowledge of the effects of huntingtin silencing outside the central nervous system is crucial to understanding the safety and efficacy of huntingtin lowering treatments.

One peripheral organ of interest is the liver. HD patients often have metabolic symptoms, such as the inability to maintain body weight, and the liver is an important regulator of metabolic homeostasis in the body (van der Burg et al., 2011; Coffey et al., 2017; Stuwe et al., 2013). The goal of this study was to explore the molecular effects of

two huntingtin gene silencing techniques, antisense oligonucleotide (ASO) or knockout (KO), in mouse liver tissue. Using gene expression and metabolite profiling data, we found several transcriptional changes induced by the huntingtin (*Htt*) gene silencing techniques, but relatively few metabolite changes. Similar functional processes were altered by the ASO and KO silencing methods.

A.2 Results and Discussion

A.2.1 Generation of mice cohorts

Two cohorts of mice were grown to understand the effects of huntingtin silencing. Both cohorts were derived from female C57Bl/6J mice and were grown for 10 months. Liver tissue from both mice cohorts at 10 months of age was harvested for RNA-Seq and untargeted polar metabolite profiling.

The first cohort, termed the ASO cohort, contains heterozygous $Htt^{Q111/+}$ and wild-type $Htt^{+/+}$ mice. The heterozygous $Htt^{Q111/+}$ mice have a mutated copy of the huntingtin gene, with an expanded polyglutamine tract of 111 repeats. In contrast, the wild-type $Htt^{+/+}$ mice have normal copies of the huntingtin gene. The case group of mice within each genotype in this ASO cohort was treated with an *Htt* ASO and the control group within each genotype was treated with either a control ASO or saline vehicle. The *Htt* ASO in the case group leads to approximately 64% knockdown of *Htt* expression in liver (Coffey et al., 2017). The complete ASO cohort contained samples from 36 mice, with 6 samples representing each the following conditions: $Htt^{+/+}$ mice with *Htt* ASO treatment, $Htt^{+/+}$ mice with control ASO treatment, $Htt^{+/+}$ mice with saline treatment, $Htt^{Q111/+}$ mice with *Htt* ASO treatment, $Htt^{Q111/+}$ mice with control ASO treatment, $Htt^{Q111/+}$ mice with saline treatment. The mice treated with control ASO or saline were grouped as one large control group, as there were no differences between these mice (Coffey et al., 2017).

The second cohort of mice, termed the KO cohort, contains only wild-type $Htt^{+/+}$ mice. The case group of mice in this KO cohort had hepatocyte-specific *Htt* knockout, generated using tissue-specific albumin-Cre drivers (Jeff Carroll, unpublished work). These mice only lack *Htt* in the hepatocytes of the liver. Because *Htt* is an essential gene, only tissue-specific knockouts are viable. These knockout mice have complete

knockdown of *Htt* in hepatocytes. The control group of mice in the KO cohort had no alteration in *Htt* expression. The complete KO cohort contained samples from 12 mice, with 6 samples representing each of the following conditions: $Htt^{+/+}$ mice with hepatocyte-specific *Htt* KO, $Htt^{+/+}$ control mice.

A.2.2 Transcriptional effects of *Htt* ASO and KO conditions

The levels of 20,067 and 20,511 genes were measured in the liver tissue from the ASO and KO cohorts, respectively. To reveal similarities between the samples, we used dimensionality reduction techniques, such as PCA, on the RNA-Seq data for each cohort (Figure A-1). We found that in the ASO cohort, the samples primarily clustered by the case and control groups. The genotype of the samples did also influence the distribution of the samples, but this effect is secondary to the treatment effect (Figure A-1A). Similarly, the samples in the KO cohort clustered based on the case and control groups (Figure A-1B). Most of the variance between the samples is likely associated with inter-mouse differences that are unrelated to genotype or treatment.

In both cohorts, *Htt* silencing lead to several transcriptional changes. In the ASO cohort, there were 3,518 significantly differentially expressed genes (FDR-adjusted p-value < 0.05) between the case and control wild-type $Htt^{+/+}$ samples. Between the case and control heterozygous $Htt^{Q111/+}$ samples, there were 3,671 significantly differentially expressed genes. In the KO cohort, there were 2,147 significantly differentially expressed genes between the case and control samples. When the differentially expressed genes were compared, we found that the ASO cohort has a genotype-specific transcriptional response (Figure A-2A). There is a strong overlap between the differentially expressed genes in the KO cohort and the differentially expressed genes in the wild-type $Htt^{+/+}$ samples of the ASO cohort. Similarly, there is a strong overlap between the differentially expressed genes in the KO cohort and the differentially expressed genes in the heterozygous $Htt^{Q111/+}$ samples of the ASO cohort. However, there is no overlap between the differentially expressed genes in the wild-type $Htt^{+/+}$ samples of the ASO cohort and the differentially expressed genes in the heterozygous $Htt^{Q111/+}$ samples of the ASO cohort.

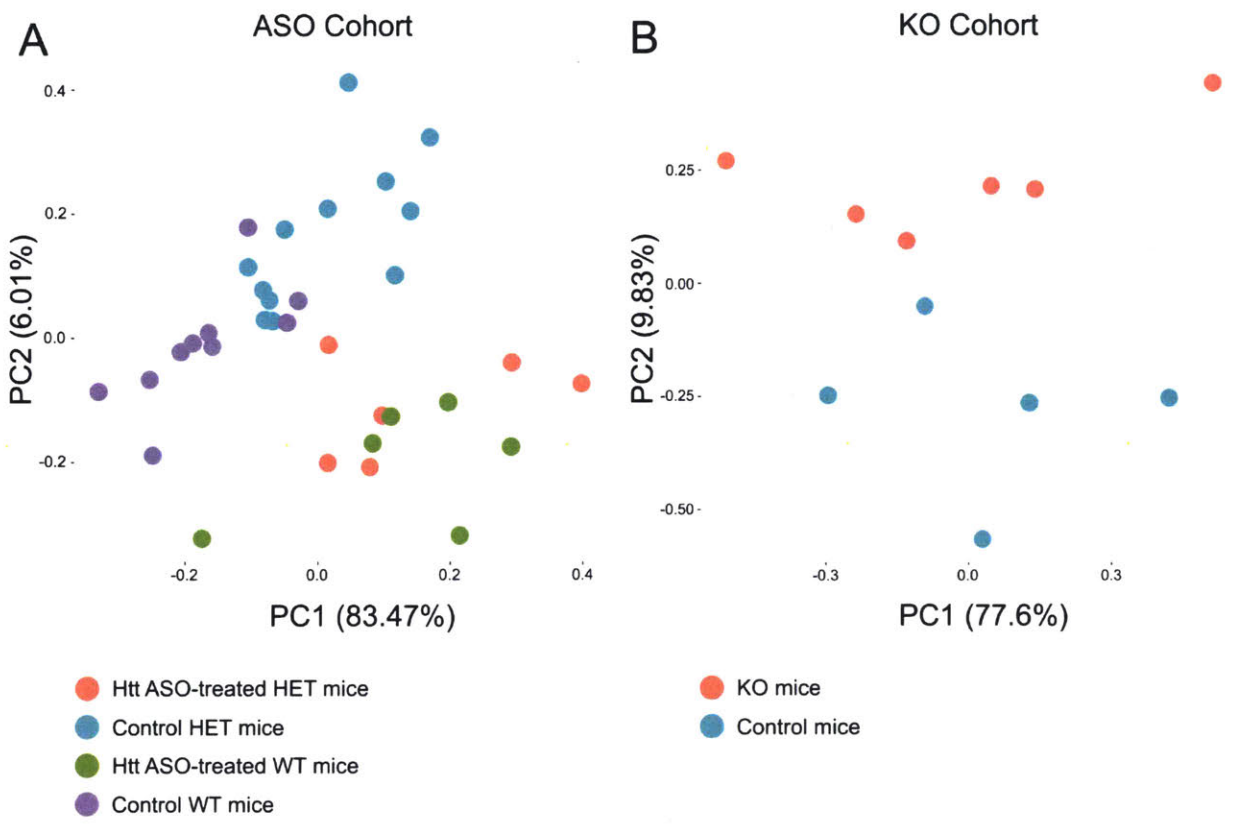

Figure A-1 *Htt* Silencing Effects on Transcription

(A) PCA plot showing the gene expression data from the ASO cohort. WT and HET refer to the wild-type or heterozygous genotype of the mice, respectively.

(B) PCA plot showing the gene expression data from the KO cohort. KO mice are those with the hepatocyte-specific *Htt* knockout.

A.2.3 Metabolomic effects of *Htt* ASO and KO conditions

In the metabolite profiling data from the ASO cohort, 1,468 untargeted metabolites were measured and passed quality control. In the metabolite profiling data from the KO cohort, 4,471 untargeted metabolites were measured and passed quality control. There were no common metabolites measured in both cohorts. The metabolomic data from both cohorts had many missing values, which could be due to mass spectrometer detection or ionization limitations. To overcome this issue of missing data, only metabolites with at least three samples per condition in each cohort were considered for further analysis. Due to presence of missing data, we performed

hierarchical clustering instead of PCA to cluster the samples. Like the RNA-Seq data, the metabolite data clusters primarily by the case and control groups in both cohorts (Figure A-3A). In the ASO cohort, the genotype of the mice did not have a strong effect on the clustering of the samples. However, there is still variance between the samples that is likely explained by inter-mouse differences that are unrelated to genotype or treatment.

In both cohorts, *Htt* silencing lead to few metabolomic changes. In the ASO cohort, there were 201 and 56 significantly differential metabolites (FDR-adjusted p-value < 0.1) between the case and control samples from wild-type $Htt^{+/+}$ and heterozygous $Htt^{Q111/+}$ mice, respectively. Unlike the transcriptomic data, the metabolite changes are not genotype-specific, as 32 metabolites were changed in both the wild-type and heterozygous samples (Figure A-2B). In the KO cohort, there were 88 significantly different metabolites in the comparison of the case and control samples.

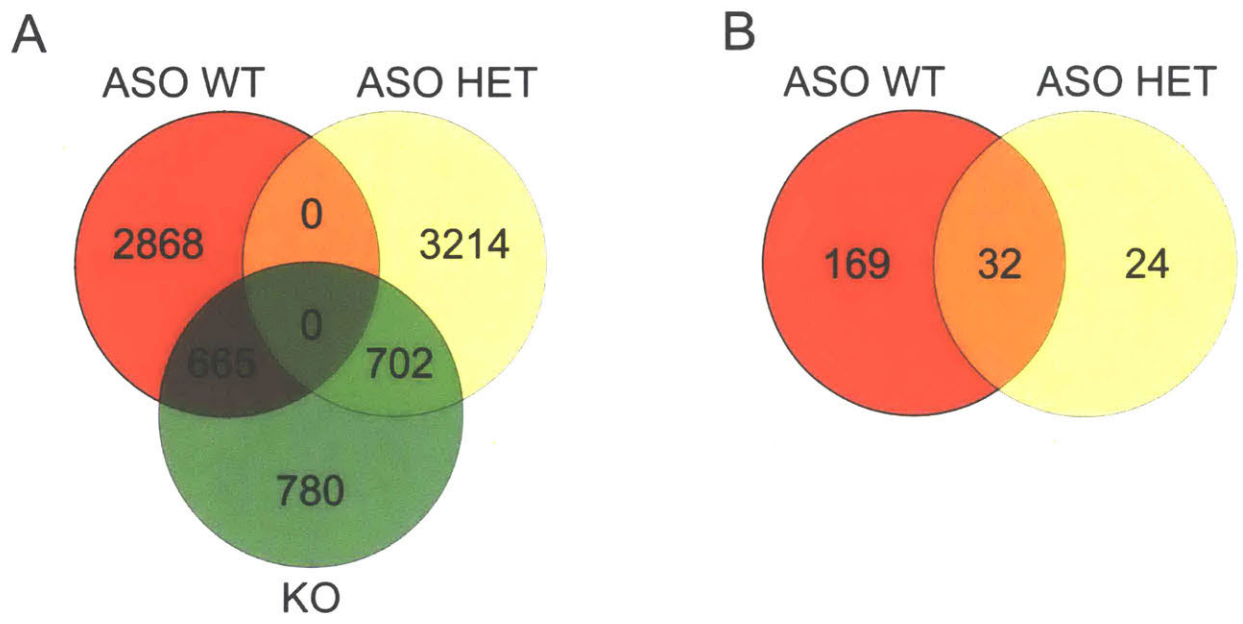

Figure A-2 Genotype-Specific Effects on Gene Expression

(A) Venn diagram showing the overlaps between the lists of significantly differentially expressed genes between cases and controls for each group of mice. ASO WT = wild-type $Htt^{+/+}$ mice from the ASO cohort; ASO HET = heterozygous $Htt^{Q111/+}$ mice from the ASO cohort; KO = mice from the KO cohort.

(B) Venn diagram showing the overlaps between the lists of significantly differentially expressed metabolites for each ASO genotype. ASO WT = wild-type $Htt^{+/+}$ mice from the ASO cohort; ASO HET = heterozygous $Htt^{Q111/+}$ mice from the ASO cohort.

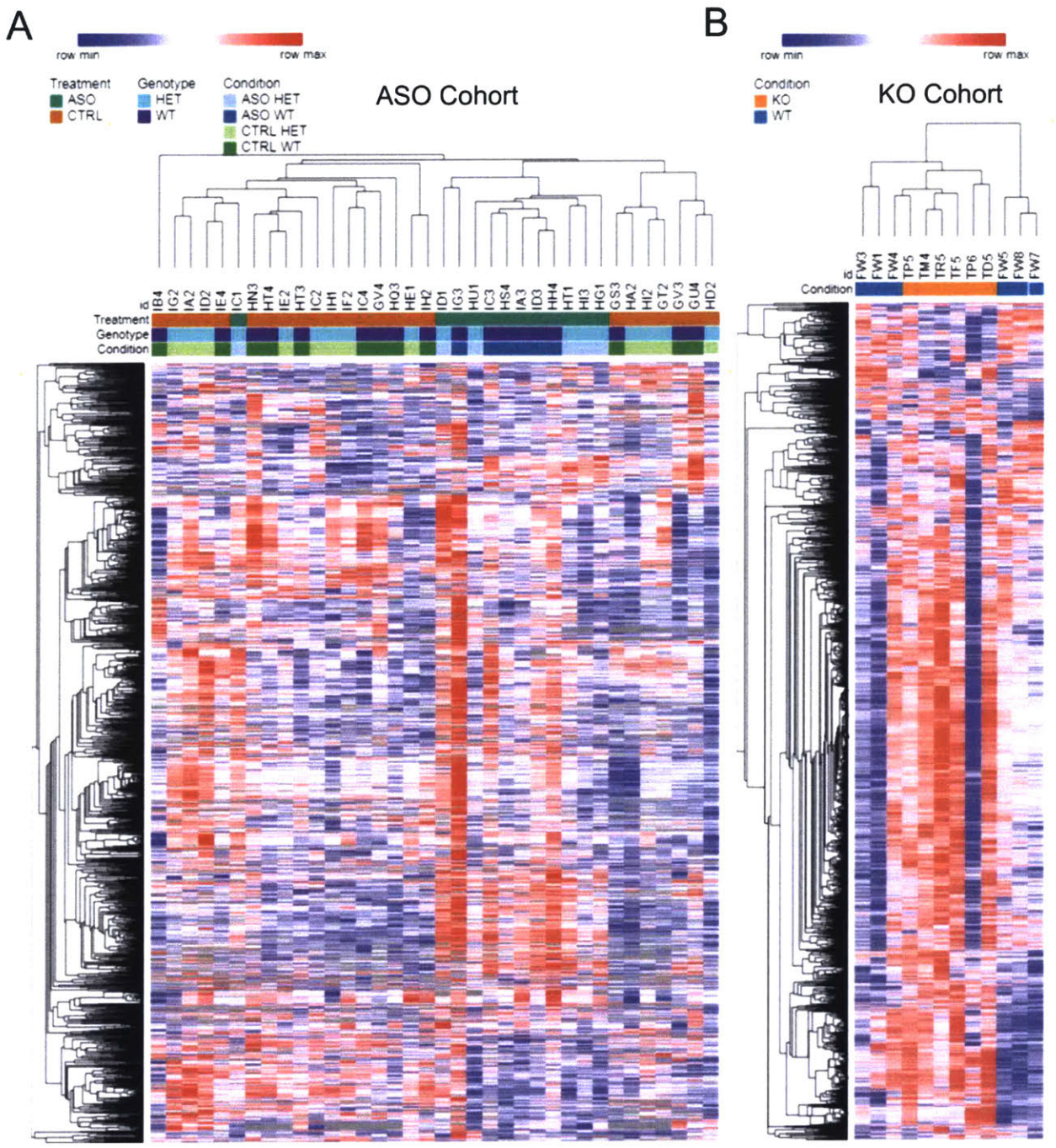

Figure A-3 *Htt* Silencing Effects on Metabolites

(A) Heatmap showing the hierarchical clustering of the metabolite data from the ASO cohort. ASO = *Htt* ASO treatment; CTRL = control treatment; WT = wild-type $Htt^{+/+}$ genotype; HET = heterozygous $Htt^{Q111/+}$ genotype.

(B) Heatmap showing the hierarchical clustering of the metabolite data from the KO cohort. KO = hepatocyte-specific *Htt* knockout, WT = control.

A.2.4 Similar cellular processes are affected by *Htt* ASO and KO silencing

Pathway enrichment analysis for the differentially expressed genes and metabolites implicated similar cellular processes altered by each *Htt* silencing condition (Figure A-4). Most of the shared processes were related to immune system and fatty acid oxidation pathways. There were some condition-specific alterations in pathways such as the electron transport chain, steroid metabolism, endocytosis, localization, and autophagy. Though these processes have been implicated in Huntington's Disease, the mechanisms by which *Htt* silencing affects these processes remains to be discovered (Martin et al., 2015; Schulte and Littleton, 2011; Zuccato et al., 2010).

To understand the connections between the transcriptomic and the metabolomic data, we performed network analysis using three different inputs. First, we generated networks with only differential metabolites as input. Because the metabolite identities are ambiguous, they were first mapped to known metabolites. However, there were very few metabolites that matched those in the PIUMet database. As a result, the networks with metabolites alone had very few nodes with associated data. Next, we leveraged publicly available DNAse-Seq data to predict transcription factors that could be regulating the observed differentially expressed genes. In these networks, the transcription factors cluster separately from the metabolites and do not provide additional useful information. The third type of network we generated was using the metabolites and differentially expressed genes as inputs. Here, the genes dominate the network and give similar pathway enrichment to the RNA-Seq pathway enrichment alone. The network analysis of this data is limited by the few differentially expressed metabolites with known interactions.

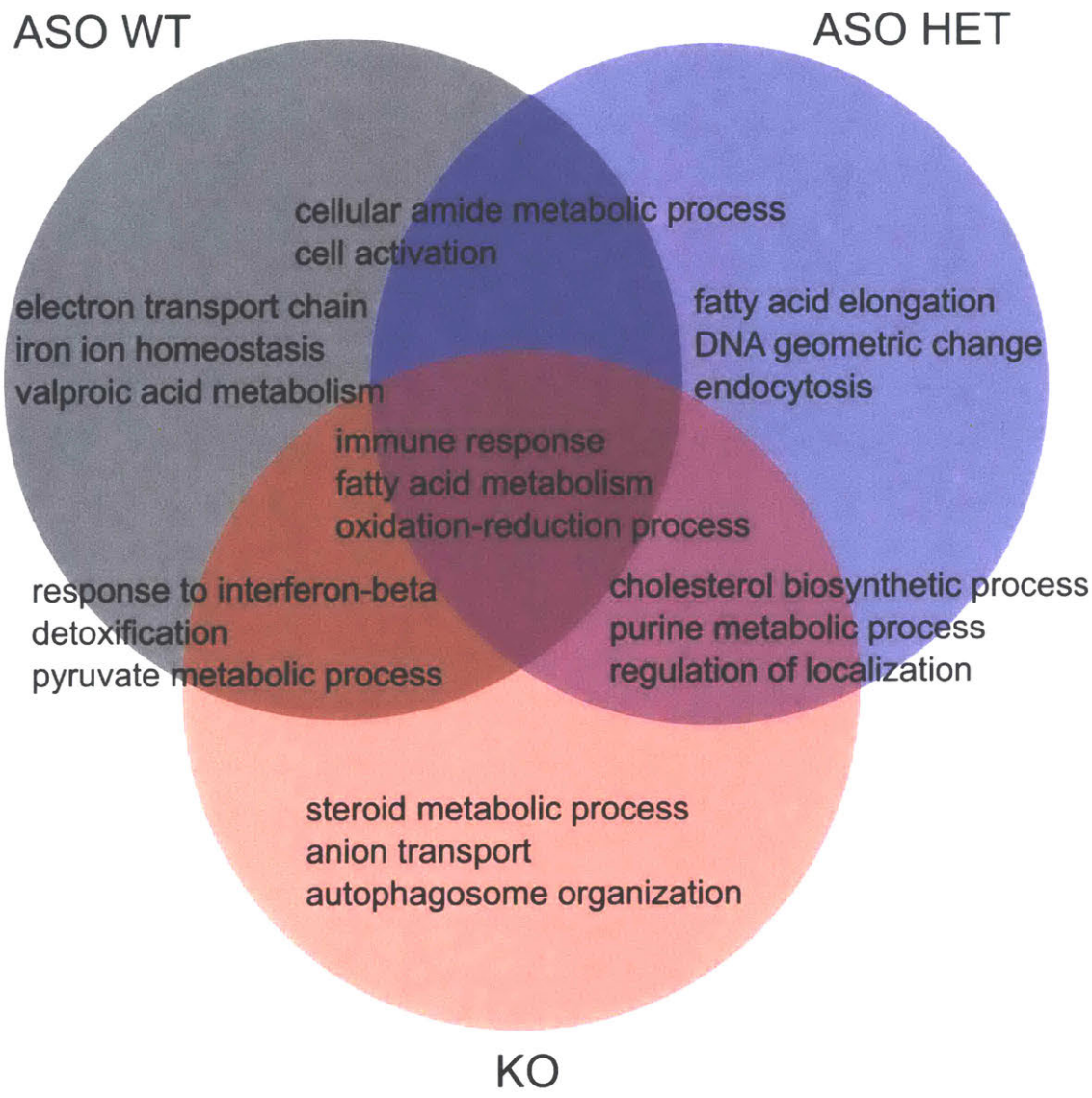

Figure A-4 *Htt* Silencing Methods Affect Similar Functional Processes

Venn diagram showing the overlap between the enriched pathways for each group of mice. The pathways were determined for each group of mice by using the differentially expressed genes and metabolites between cases and controls. ASO WT = wild-type $Htt^{+/+}$ mice from the ASO cohort; ASO HET = heterozygous $Htt^{Q111/+}$ mice from the ASO cohort; KO = mice from the KO cohort.

The lack of a strong metabolomic effect could be due to many reasons. First, previous studies report a modest phenotype difference between *Htt* ASO-treated mice and control mice. For example, the *Htt* silenced mice from this study had only a 5%

decrease in body mass compared to the control mice (Coffey et al., 2017). Perhaps the differences at the metabolite level are too small to detect with the cohort size and the type of data collected. Also, it is known that peripheral silencing does not change the severity of HD-relevant phenotypes in the striatum of *Htt*$^{Q111/+}$ mice (Coffey et al., 2017). It is reasonable that the silencing effects in the liver or hepatocytes specifically could be masked by unknown interactions and compensatory effects with other organs or cell types in the mice. Overall, there are few functional differences between the *Htt* silencing approaches in the liver. Though there were several transcriptomic changes, many of which were genotype-specific, the cellular processes associated with those changes implicated the same biological pathways of immune system and fatty acid oxidation processes. Future studies could investigate the role of *Htt* in these processes and compare the effects of silencing in other peripheral tissues.

A.3 Methods

Mouse Models of Huntingtin Silencing

Female C57Bl/6J heterozygous *Htt*$^{Q111/+}$ and wild-type *Htt*$^{+/+}$ mice were acquired from the Jackson Laboratories (Bar Harbor, ME) and grown and treated with *Htt* and control ASOs as previously described (Coffey et al., 2017). Using tissue-specific albumin-Cre drivers in *Htt*$^{+/+}$ mice, hepatocyte-specific *Htt* knockout mice were generated (Jeff Carroll, unpublished). Liver tissue was harvested at 10 months of age from both mice cohorts for RNA-Seq and untargeted polar metabolite profiling.

Differentially Expressed Genes

Adapter sequences were trimmed from sequencing reads using Trim Galore v0.4.2 (https://github.com/FelixKrueger/TrimGalore). Paired-end reads were aligned to the mm10 UCSC reference genome (http://genome.ucsc.edu/) and quantified using TopHat2 (Kim et al., 2013). For differential expression analysis, one wild-type control sample was removed from both cohorts due to technical sample issues. DESeq2 was used to find differentially expressed genes for each *Htt* silencing condition compared to its relevant control (Love et al., 2014). The differentially expressed genes were filtered using a Benjamini-Hochberg corrected p-value threshold of 0.05.

Differentially Expressed Metabolites

Metabolite quantification in positive and negative ionization mode was filtered using the following quality control checks: removed any values with metabolite intensity less than 100; removed any metabolites where the 10 quality control injections had a CV greater than or equal to 0.25. The data was then normalized by protein level per sample. Metabolites with abundance measurements for at least three replicates per condition were then log2 normalized and analyzed using limma (Ritchie et al., 2015). Differentially expressed metabolites were filtered using a Benjamini-Hochberg corrected p-value threshold of 0.1. Untargeted metabolite m/z peaks were matched to known metabolites using PIUMet, with a metabolite database compiled using HMDBv4.0 and Recon3D (Brunk et al., 2018; Pirhaji et al., 2016; Wishart et al., 2018).

Dimensionality Reduction and Clustering

We displayed the gene expression data as PCA plots using the stats package in R (R Core Team, 2017). We hierarchically clustered the metabolite profiling data using the Spearman rank correlation and created heatmaps using Morpheus (https://software.broadinstitute.org/morpheus).

Pathway Enrichment

Enrichment analyses of the differential genes and network proteins were performed using GOrilla with a background set of all genes measured or all proteins present in the interactome, respectively (Eden et al., 2009). Enrichment analyses of the differential metabolites were performed using IMPaLA with a background set of all metabolites measured (Kamburov et al., 2011). A Benjamini-Hochberg correct p-value threshold of 0.05 was applied to assign significant to the pathway enrichment terms.

Transcription Factor Prediction

DNAse-Seq data for three male wild-type 8-week adult C57Bl/6J mice were downloaded from ENCODE (https://www.encodeproject.org/) with the following identifiers : ENCFF001PRR, ENCFF001PRT, ENCFF001PRS (Davis et al., 2018; Dunham et al., 2012). Adapter sequences were trimmed from sequencing reads using

Trim Galore v0.4.2 (https://github.com/FelixKrueger/TrimGalore). Bowtie2 and samtools v1.3 were used to sort and index the reads, as well as remove mitochondrial DNA (Langmead and Salzberg, 2012; Li et al., 2009). Peaks were called using MACS2 (Zhang et al., 2008). Motif analysis was used to predict transcription factors that could be regulating the differentially expressed genes. Motifs were annotated to the mm10 UCSC reference genome (http://genome.ucsc.edu/) using the CIS-BP database (Waterston et al., 2002; Weirauch et al., 2014). A hypergeometric test was used for each transcription factor to find those with motifs in regions intersecting DNAse-Seq peaks and within 2kb of differentially expression genes for a given condition. A Benjamini-Hochberg corrected p-value threshold of 1E-5 was applied to assign significance to transcription factor predictions.

Network Analysis

Differential m/z metabolite peaks, predicted transcription factors, and differentially expressed genes for each *Htt* silencing condition compared to control were mapped onto the interactome, comprised of physical interactions between proteins (iRefIndex v14), proteins and metabolites (HMDBv4.0, Recon3D), and m/z peaks and matched metabolites (PIUMet) (Brunk et al., 2018; Pirhaji et al., 2016; Razick et al., 2008; Wishart et al., 2018). The Prize-Collecting Steiner Forest (PCSF) algorithm was applied using Omics Integrator 2 to find the set of highly relevant pathways associated with each compound treatment (Tuncbag et al., 2016). PCSF was run 100 times with random noise on the edges for robustness measurements and random input sets for specificity measurements. The optimal network solution was filtered by those nodes with at least 20% robustness and specificity.

Networks were visualized in Cytoscape (Shannon et al., 2003). In each network, the nodes are a combination of proteins, transcription factors, or metabolites. The integration of the RNA-Seq and DNAse-Seq data provided transcription factor predictions. Some networks were created with RNA-Seq data as input, but these were treated cautiously as the interactome is based on protein interactions, not gene interactions.

A.4 References

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. *36*, 272–281.

van der Burg, J.M.M., Winqvist, A., Aziz, N.A., Maat-Schieman, M.L.C., Roos, R.A.C., Bates, G.P., Brundin, P., Björkqvist, M., and Wierup, N. (2011). Gastrointestinal dysfunction contributes to weight loss in Huntington's disease mice. Neurobiol. Dis. *44*, 1–8.

Coffey, S.R., Bragg, R.M., Minnig, S., Ament, S.A., Cantle, J.P., Glickenhaus, A., Shelnut, D., Carrillo, J.M., Shuttleworth, D.D., Rodier, J.-A., et al. (2017). Peripheral huntingtin silencing does not ameliorate central signs of disease in the B6.HttQ111/+ mouse model of Huntington's disease. PLoS One *12*, e0175968.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Dunham, I., Bernstein, B.E., Birney, E., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Kaemmerer, W., and Grondin, R. (2019). The effects of huntingtin-lowering: what do we know so far? Degener. Neurol. Neuromuscul. Dis. *9*, 3–17.

Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R., and Keun, H.C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics *27*, 2917–2918.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

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

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

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Martin, D.D.O., Ladha, S., Ehrnhoefer, D.E., and Hayden, M.R. (2015). Autophagy in Huntington disease and huntingtin in autophagy. Trends Neurosci. *38*, 26–35.

Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C.B., White, F.M., Saghatelian, A., and Fraenkel, E. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. Nat. Methods *13*, 770–776.

R Core Team (2017). R: A language and environment for statistical computing. http://www.R-project.org/.

Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: A consolidated

protein interaction database with provenance. BMC Bioinformatics *9*, 405.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
Schulte, J., and Littleton, J.T. (2011). The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. Curr. Trends Neurol. *5*, 65–78.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.
Stuwe, S.H., Goetze, O., Lukas, C., Klotz, P., Hoffmann, R., Banasch, M., Orth, M., Schmidt, W.E., Gold, R., and Saft, C. (2013). Hepatic mitochondrial dysfunction in manifest and premanifest Huntington disease. Neurology *80*, 743–746.
Tabrizi, S.J., Ghosh, R., and Leavitt, B.R. (2019). Huntingtin Lowering Strategies for Disease Modification in Huntington's Disease. Neuron *101*, 801–819.
Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput. Biol. *12*, e1004879.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.
Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell *158*, 1431–1443.
Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., et al. (2018). HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Res. *46*, D1074-82.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.
Zuccato, C., Valenza, M., and Cattaneo, E. (2010). Molecular Mechanisms and Potential Therapeutical Targets in Huntington ' s Disease. Physiol Rev *90*, 905–981.

Appendix B: An integrated multi-omic analysis in iPSC-derived motor neurons from *C9ORF72* ALS patients

This work is being prepared for publication.

The NeuroLINCS Consortium*. An integrated multi-omic analysis in iPSC-derived motor neurons from *C9ORF72* ALS patients.

*Relevant groups in the NeuroLINCS Consortium:
Epigenomics: Pamela Milani, Miriam Adam, Brook T Wassie, Ernest Fraenkel.
Integrative Analysis and Computational Modeling: Jonathan Li, Renan Escalante-Chong, Alex Lenail, Karen Sachs, Ryan Lim, Julia Kaye, Natasha L Patel-Murray, Divya Ramamoorthy, Steven Finkbeiner, Leslie M Thompson, Ernest Fraenkel.

As part of this work, I would like to acknowledge ALS patients and their families for their essential contributions to this research.

My contributions:
I performed preliminary data analysis and network modeling in the early stages of this collaborative project. I also worked on project management tasks, such as website design, metadata, and data releases.

To conserve space, I have left out the supplemental figures and tables. The final supplementary information can be found in the publication.

B.1 Abstract

Systematically mapping molecular changes occurring early in neurodegenerative diseases prior to symptom onset could dramatically accelerate and broaden therapeutic strategies. The NeuroLINCS consortium produced a detailed molecular characterization of motor neurons from induced pluripotent stem cells (iPSCs) derived from patients with amyotrophic lateral sclerosis (ALS) who carried hexanucleotide expansions in *C9ORF72* - the most common known cause of ALS. There were no significant differences in iPSC or motor neuron generation between ALS and control subject lines. Searching for early molecular differences, we characterized cellular states through whole genome sequencing, ATAC-seq, RNA-seq, and data-independent acquisition mass-spectrometry (DIA-MS) proteomics. Several pathways, including biological adhesion and extracellular matrix organization, were altered across epigenomic, transcriptomic, and proteomic data, although few individual genes showed consistent changes. Using novel computational methods, we discovered molecular networks linking alterations across the data modalities, uncovering key transcriptional regulators. To distinguish between causal versus compensatory pathway changes induced by *C9ORF72* expansions, we tested network genes modifying ALS in a *C9ORF72* Drosophila model. This revealed causal pathways including RNA processing, transport and translation, and compensatory pathways such as DNA repair and transcriptional regulation. This new integrated NeuroLINCS data set has been posted on a data portal that allows scientists worldwide to explore, challenge, and generate new disease-related hypotheses.

B.2 Introduction

Modeling neurological diseases using induced pluripotent stem cell (iPSC) technology offers a unique platform to study the process of pathogenesis. Rather than using artificially expressed human disease genes in mice or end stage post mortem tissues from patients, the generation of new neurons and astrocytes from patient-specific cells allows for discovery of the earliest genesis of disease signatures. One neurodegenerative disease group that has been modeled extensively using iPSCs are the motor neuron disorders. Adult onset motor neuron diseases include amyotrophic

lateral sclerosis (ALS), where motor neurons degenerate late in life, inevitably leading to paralysis and asphyxiation. Genetic underpinnings have been identified in ~15% of ALS cases (Paez-Colasante et al., 2015). Of these, the most common mutation is a hexanucleotide repeat expansion in the first intronic region of *C9ORF72* which accounts for over 30% of all known genetic forms of the disease. While much is known about the mutation and the abnormal proteins that are produced by its transcripts, it is still unclear how repeats in *C9ORF72* ultimately lead to cellular dysfunction and death (Brown and Al-Chalabi, 2017).

Some of the first disease modeling studies showed that iPSCs could be generated from early onset motor neuron diseases, such as spinal muscular atrophy, and that these motor neurons exhibited disease-specific cell death in the petri dish (Ebert et al., 2009; Fuller et al., 2016; Ng et al., 2015; Nizzardo et al., 2015; Sareen et al., 2012; Vazquez-Arango et al., 2016). Interestingly, for later onset motor neuron diseases, such as ALS, iPSC models did not initially show any overt death in motor neurons (Dimos et al., 2008). However, for inherited forms of ALS, such as *C9ORF72* repeat expansions (C9), there were specific changes in neuron activity, gene expression, and cellular processes (Devlin et al., 2015; Donnelly et al., 2013; Sareen et al., 2013; Selvaraj et al., 2018; Shi et al., 2018; Wainger et al., 2014). More recently, stressors such as trophic factor withdrawal have led to ALS-specific cell death phenotypes, although it is not clear how these stressors relate to human disease onset and progression (Shi et al., 2018). In a very recent study, subsets of sporadic ALS patients also showed phenotypic changes including reduced fiber outgrowth at later time points in culture, although a comprehensive omics analysis was not performed (Fujimori et al., 2018).

These iPSC models provide a unique opportunity to examine the molecular changes that occur due to ALS causing genes in motor neurons. While post-mortem studies have provided important insights into these processes, patient samples represent a late stage of the disease, which may not exhibit molecular or cellular signatures directly associated with events that cause the disease (Delic et al., 2018; Emde et al., 2015; Paré et al., 2018; Prudencio et al., 2015; Sanfilippo et al., 2017). By contrast, cells derived from iPSCs can provide insights into the earliest stages of

neurodegeneration, opening a window into the period when therapeutics might have the greatest benefit.

The goal of the current study was to test whether using a multi-omic approach and network-based analysis would facilitate identification of pathogenic events that define C9 ALS. This more complete description of the pathogenic process would enable the discovery of new disease pathways and subsequently new drug targets. We developed an integrative approach that combined multi-omic data with functional experiments in a Drosophila model to distinguish causal and compensatory pathways involving the extracellular matrix, microtubules, and the nuclear pore complex. As part of the NIH-funded NeuroLINCS consortium, all of the data sets along with the data integration have been posted to a portal for data sharing and crowd sourcing of this unique resource http://neurolincs.org.

B.3 Results

B.3.1 Generation and characterization of iPSC lines

The control and C9-ALS iPSC lines used in this study were generated using episomal plasmid-based reprogramming methods, and all lines retained their repeat expansion mutation following reprogramming as described previously (Sareen et al., 2013). All iPSC lines maintained normal karyotypes as determined by G-band karyotyping and the identity of iPSCs and differentiated iMNs were confirmed to match the parent fibroblasts by DNA fingerprinting.

B.3.2 Whole genome sequencing shows no overt abnormalities

Whole genome sequencing (WGS) was performed on all of iPSC lines from three healthy controls and four with ALS-associated hexanucleotide repeat expansions (HRE) in *C9ORF72* previously described in detail (Sareen et al., 2013). A novel computational pipeline was used to annotate the variants in the genomes of the control and C9-ALS lines relative to reference human genomes. The number of single nucleotide polymorphisms (SNPs) was within the expected range, and there were no overt genetic abnormalities. Across all lines, we found 11,260,464 variants with 9,197,462 variants in the control lines and 8,818,235 variants in the C9-ALS lines. Thus, there was an

average of 5.4 million variants per line, which is consistent with the variation that has been previously observed in human genomes (Auton et al., 2015).

After applying annotation, we filtered for exonic functional variation (Table B-1). There were 57,910 exonic functional variants in the controls, and 12,898 were rare (less than 1%) or novel (no frequency information). There were 55,815 exonic functional variants in the C9-ALS lines, and 8,225 were rare or novel. Next, we investigated if any of the lines had genetic variants previously associated with ALS and found 3 variants in *OPTN*, *ALS2* and *DIAPH3*. Other variants in ALS-associated genes were observed, but none that were known previously to be disease-associated or causing. However, of interest, the 52i ALS line contains the *APOE*-ε4 allele (rs429358) (C130R) which is associated with an increased risk of Alzheimer's disease (Farrer et al., 1997). We next applied the American College of Medical Genetics gene criteria to identify likely pathogenic (LP) or Pathogenic (P) variants. Although a subset of these variants is in ALS genes that are listed in the ASLoD database, to our knowledge none of these variants are expected to confer risk of developing ALS (Wroe et al., 2008). Overall, WGS analysis of the patient cell lines revealed no pathogenic or likely pathogenic variants that to our knowledge are not expected to interfere with disease progression of ALS *per se*.

B.3.3 C9 phenotypic signatures in iPSC-derived motor neuron cultures

The iPSC lines were first patterned into motor neuron precursor spheres (iMPS), expanded as suspension aggregate cultures for ~5 weeks using a chopping method, and subsequently seeded to differentiate into motor neuron cultures (iMNs) for another 21 days (Figure B-1A) (Shelley et al., 2014). These iMN cultures were harvested, equally distributed in three replicate cell pellets and analyzed using the multi-omics (transcriptomics, proteomics, and ATAC-seq) assays. Both control and C9-iPSC lines gave rise to similar numbers of neurons and glia based on immunocytochemical staining for SMI32, TuJ1 (β3-tubulin), Map2a/b, GFAP, and nestin (Figure B-1B,C).

To establish if there were differences in survival between C9-ALS and control motor neurons, cells were transfected with the motor neuron morphology marker HB9-GFP and given approximately 3-4 days of recovery time (Wilson, 2005). Neurons were

then subjected to automated longitudinal imaging for 7 days twice a day. Images were collected and montaged for analysis. Fluorescent cells displaying the typical neuronal morphology, including soma along with thin axon-like processes tipped with growth cones, were tracked to determine the cumulative risk of death (Figure B-2A). We evaluated the survival for each fluorescent neuron and this revealed that there was no increased risk of death in C9-ALS lines compared to controls. In fact, we were surprised to see the contrary in that C9-ALS lines survived significantly better than control lines (Figure B-2B). Collectively, this suggests that there are no overt signs of changes to neuronal maturation or degeneration in C9-ALS lines.

B.3.4 Transcriptomic analysis reveals known and novel pathways related to C9

RNA sequencing revealed specific transcriptomic signatures associated with the C9 lines. Total RNASeq (Ribo-Zero rRNA depletion) was carried out on the distributed iMN pellets as described in methods. Statistical analysis of differential expression was analyzed using DESeq2. We found 828 differentially expressed transcripts (271 downregulated and 557 upregulated) between C9-ALS and control iMNs (FDR < 0.1), of which 704 were annotated as protein-coding in Uniprot. Exploratory analysis of gene expression levels was performed using hierarchical clustering (Figure B-3A). To begin to understand the effect of the C9 mutation on a multicellular culture, genes that were significantly different between C9 and control samples were used for Cell Type-Specific Expression Analysis (CSEA) (Bossis et al., 2005).

CSEA revealed an enrichment of cortical and motor neuron specific gene expression. Next, gene ontology (GO) analysis was conducted to determine the functional role of these genes, using GOrilla on the 704 DEGs, revealing an enrichment in extracellular matrix (ECM) and cell adhesion terms which included: ECM disassembly, ECM organization, collagen binding, and focal adhesion (Figure B-3C). A previous study of C9 iPSC-derived motor neurons showed dysregulation of 66 genes between 4 ALS and 4 control samples with a fold change of > 2 and p-value < 0.05 (Sareen et al., 2013). Of those 66 genes, 8 genes overlapped with the 828 DEGs from our study, although in different directions. Even with this small number of overlapping

genes, GO enrichment analysis revealed an enrichment for extracellular regions in the 66 genes, similar to our analysis.

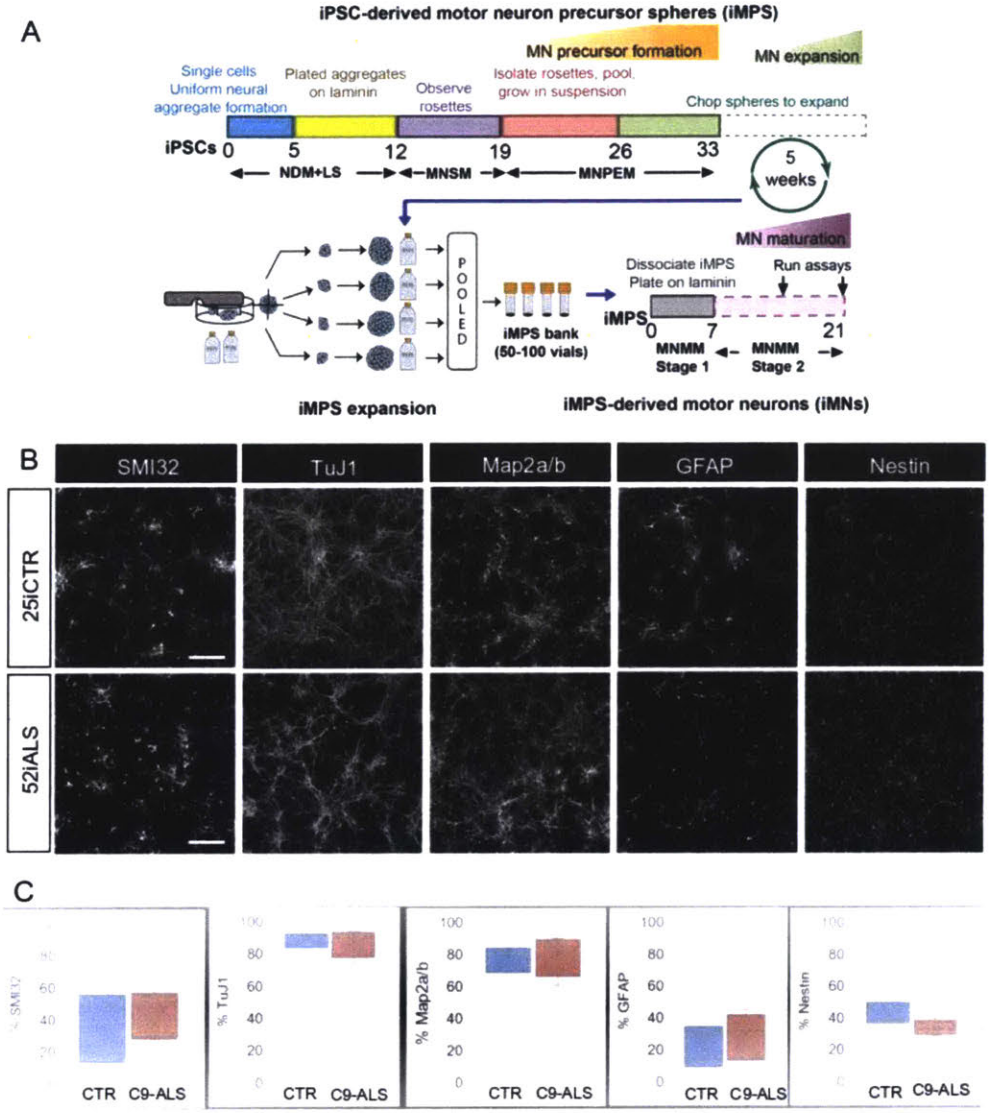

Figure B-1 (A) Schematic of protocol for iPSC differentiation into motor neuron cultures used by NeuroLINCS for transcriptomics, proteomics and ATAC-seq assays. The iPSC-derived motor neuron precursor spheres (iMPS) were dissociated into single cells from C9-ALS and healthy patient iPSC lines and plated on laminin substrate to differentiate further into motor neuron (iMN) cultures over 21 days. (B) Representative images of iMNs from control (25iCTR) and C9-ALS (52iALS) iMPS shows consistent distribution of neural cell populations marked by SMI32, TuJ1, Map2a/b, GFAP and nestin. Scale bars are 50μm. (C) Box plots quantifying levels of SMI32, Tuj1, GFAP, nestin and Map2a/b in control and C9-ALS iMN cultures.

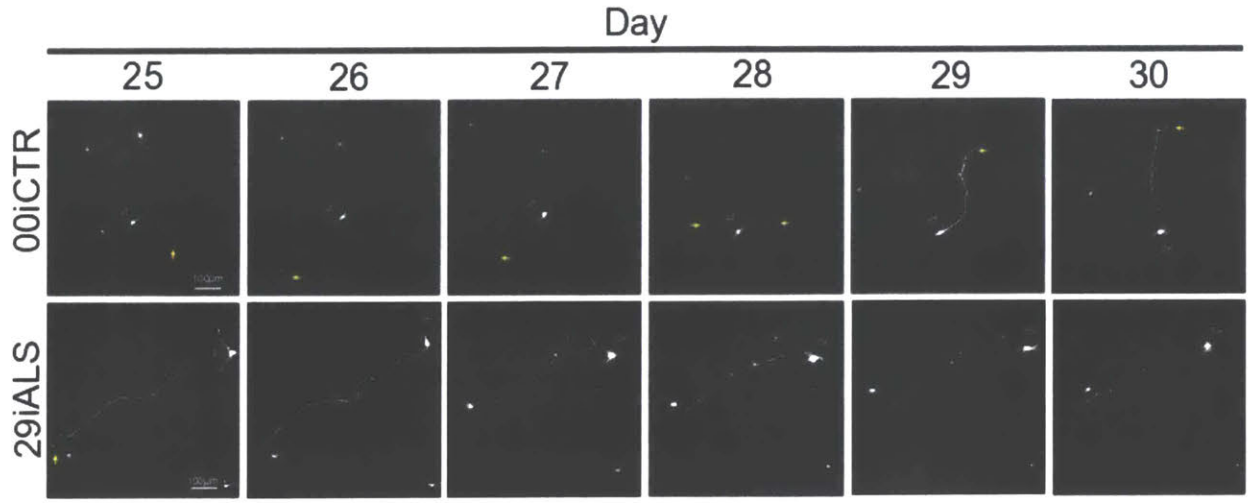

Figure B-2 (A) Representative images of iMNs over time, differentiated from control and C9-ALS lines. These cells were imaged every 12 hours over 7 days from differentiation day 25 to day 31 with a fully automated robotic microscopy system. iMNs expressed the fluorescent reporter HB9-GFP70. Motor neurons (cell bodies are indicated by hollow arrows, top and bottom rows) are seen exhibiting 1-3 processes tipped by structures resembling growth cones (indicated by golden arrow) from day 25 onwards followed by degeneration and cell death at later time points for both 00iCTR and 29iALS lines. Scale bars are 100μm.

To identify potential regulators controlling the differential expression of these ECM related genes, Ingenuity pathway analysis (IPA) upstream regulator analysis was conducted. Some of the top predicted regulators identified include SMADs (transforming growth factor beta (TGFβ) signaling), mitogen-activated protein kinase 1 (ERK), and nuclear factor kappa B (NFκB). Network-based analysis of upstream regulators and gene targets showed a TGFβ, AP-1 transcription factor subunit (AP1), erb-b2 receptor tyrosine kinase 2 (ERBB2), plasminogen activator, urokinase receptor (PLAUR), and neuregulin 1 (NRG1) network that were again predicted to regulate many of the ECM and cell adhesion related DEGs. Notably, NRG1 was identified as a major hub gene that could regulate other upstream regulators and directly regulate *ACTIN* and *INTEGRIN* expression, each of which was upregulated in the ALS iMNs. Matrix metalloproteinases (*MMPs*) were significantly dysregulated, in all cases showing

increased expression, and were downstream of the NRG1 hub (Figure B-4C). We further investigated dysregulation of these *MMPs* and found that their corresponding substrates (e.g. *LAMININs*, *COLLAGENs*) were also upregulated. These data indicate a potential role for NRG1 in the dysregulation of ECM and cell adhesion-related genes in ALS iMNs, as suggested previously in mouse models of ALS (Song et al., 2012).

Further analysis of the transcriptomic data focused on differential exon usage and alternative splicing. These analyses were conducted using DEXSeq and MATS, respectively (Anders et al., 2015; Shen et al., 2014). Analysis of the alternative splicing events found in the ALS iMNs compared to controls shows a high percentage of exon skipping (ES, 57%) and intron retention (RI, 26%). This same pattern was previously identified as enriched in studies using human familial ALS and sporadic ALS patient tissue (Prudencio et al., 2015).

B.3.5 Proteomics shows ECM and mRNA processing dominate protein changes

A sample specific library using DDA based acquisition files was compiled and DIA-MS samples were run against the peptide library. Data quality was assessed by MS1 and MS2 total ion current, normalized protein intensity distribution, number of unique and shared hits identified, and correlation between ALS and control lines. Using this method, we were able to identify 3,844 unambiguous proteins based on 23,436 unique peptides. MAP DIA software was then used to determine relative peptide and protein amounts within the samples, as well as log2FC between C9 ALS and control using transition level data. Using a 1% FDR, 95% confidence interval and 0.6 abs(log2FC) cutoff, a final list of 924 differentially expressed proteins was obtained. Hierarchical clustering of differential protein intensity values showed similar groupings between biological replicates for ALS and control samples, as seen for RNA-seq and ATAC-seq (Figure B-3A). Interestingly, unbiased analysis of all measured proteins resulted in separation between control and ALS groups.

A small subset of the differentially expressed proteins (6.8%) had overlap with both the ATAC-seq and RNA-Seq differentially expressed genes (Figure B-3B), specifically 68 common differentially expressed genes/proteins (45 between RNA and protein and 23 between all omics data sets) (Figure B-3C). The fold change values of

these overlapping terms have a correlation $R^2 = 0.76$, suggesting that most of the differentially expressed terms that are common have concordant fold change values and directionality. Of these common proteins, downregulated proteins (13) did not yield any GO enrichment terms. Common upregulated proteins/genes (55) show enrichment in extracellular matrix terms.

The role of the extracellular matrix is further supported by the analysis of the 856 differentially expressed proteins that did not overlap with differentially expressed genes. Of these, 183 proteins were upregulated and enriched for extracellular matrix proteins, similar to the transcriptomic analysis. In addition, network-based analysis of all differentially expressed proteins (924) by IPA revealed predicted upstream regulators, including TGFβ and SMAD4, which in turn regulate many of the extracellular matrix proteins identified in the differential protein analysis and integrated omics.

The remaining unique subset of proteins (674 differentially expressed proteins) were downregulated and showed enrichment for poly(A) RNA binding, RNA binding, RNA and mRNA splicing. Additionally, IPA analysis of the differential proteins (924) shows predicted inhibition of RNA/mRNA splicing based on downregulation of proteins associated with this pathway. Lastly, proteins associated with alternative splicing of mRNA are dysregulated, with most of these proteins decreasing in ALS. Taken together, this could imply that these downregulated proteins are associated with altered exon usage and alternative splicing in ALS found in the transcriptomic analysis.

B.3.6 Epigenetic changes due to C9 expression seen with ATAC-Seq

We sought to study the accessible chromatin landscape in C9 patients and controls. The density of transposase Tn5 cleavage fragments provides a continuous measurement of chromatin accessibility via ATAC-seq. Analysis of the open chromatin data identified 128,299 peaks that were active in 2 or more ALS or control samples. Approximately 14% (18,407) of accessible regions localize to gene promoters as defined by GENCODE (Harrow et al., 2012); 27% (34,543) lie within 2.5 kb of a TSS. Nearly half of the peaks lie in intronic regions, while about a third lie between genes.

To study alterations in chromatin accessibility in the disease state, we identified and characterized peaks with significantly changed accessibility between C9 and control

samples. Roughly 12% (15,814 peaks; FDR < 0.1) of all peaks were found to be differentially open, of which approximately one half (7,937) were less accessible in C9 samples. Hierarchical clustering of differentially open regions revealed similar groupings of patient samples as in RNA-Seq and proteomics (Figure B-3A). Correlation coefficients were 0.46 for RNA and ATAC and 0.13 for protein and ATAC, with both comparisons indicating same direction. Differentially accessible peaks were biased away from regions near TSSs, with only 5.0% (783) annotated to promoters. Examples of changing chromatin accessibility in ALS versus control lines can be seen in data files. Next, we sought to answer whether chromatin changes are influencing broad categories of genes by assigning each peak to its nearest RefSeq gene TSS within 50kb. 2,345 genes were associated with more ALS peaks than control and were enriched for signaling and calcium ion binding GO terms. Conversely, 2,617 genes were associated with more control peaks than ALS and were enriched for terms such as neuron development and axon guidance. Overall, ATAC-Seq identified many regulatory changes that were consistently different across ALS and control lines. In the data integration section, we analyzed how these changes correspond to changes in RNA-seq to understand differences in gene regulation between disease and control states.

B.3.7 Comparison of RNA-Seq, proteomics, and ATAC-Seq experiments

We sought to characterize the similarities and differences between the genomics, RNA-Seq, proteomics, and ATAC-seq experiments. We first examined the overlap of the RNA, protein, and epigenomics assays. Each differentially open region was assigned the nearest protein coding gene (up to a limit of 50kb from the TSS). The sets of genes and proteins detected by each assay all showed a modest increase in overlap compared to what would have been expected by chance. For example, approximately 7% of the proteins that differed between ALS and control samples were also differentially expressed in the RNA-Seq data (p-value = 1.92E-14). A higher fraction of genes that differed in RNA expression also showed changes in ATAC-seq (38%; p-value = 1.86E-14) and 14% of the proteins that differed between ALS and control samples were also differential ATAC-seq genes (p-value = 0.056). All three assays were enriched for similar biological processes. For instance, when we compared the top GO

terms from each experiment, we found that all were enriched for adhesion and extracellular matrix processes (Figure B-3C).

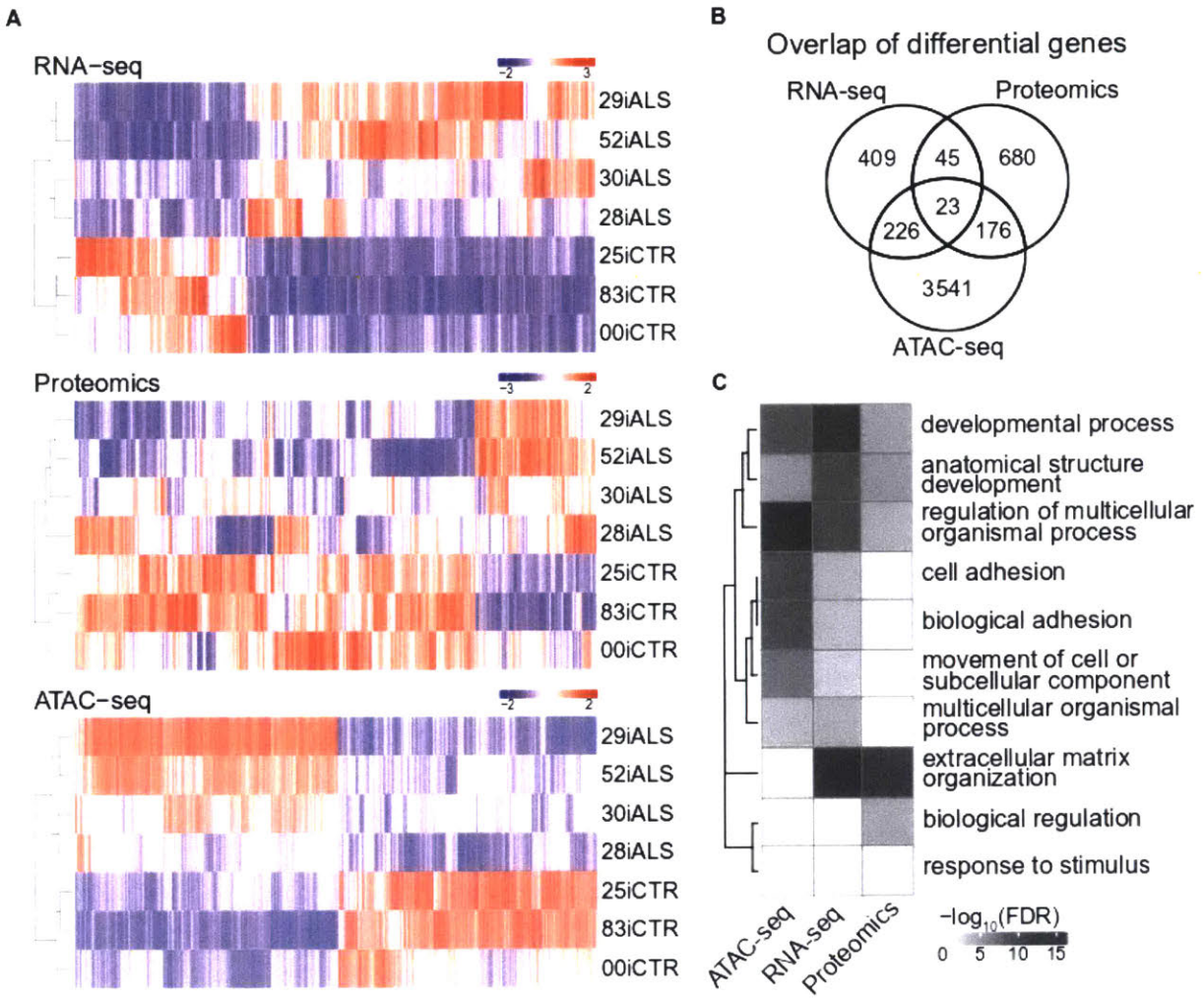

Figure B-3 (A) Hierarchical clustering of RNA-Seq, Proteomics, and ATAC-seq signals normalized by z-score. (B) Venn diagram of differential genes or proteins from each assay. Each differential ATAC-seq peak was assigned the nearest protein coding gene (up to a limit of 50kb from the TSS). (C) Top GO term enrichments for each assay reveal common biological processes.

B.3.8 WGA and RNA-Seq data integration mitigate eQTL effects on C9 dysregulation

Our analysis of the control and ALS lines revealed genomic variants in loci other than the *C9ORF72* locus that could potentially contribute to the line-specific differences in the RNA-Seq and proteomic data. Therefore, we evaluated whether any of the genetic coding variants outside the *C9ORF72* locus were disproportionately present in C9-ALS lines compared with controls to better identify differences specifically attributable to ALS-associated HRE in *C9ORF72*. For example, we observed that a missense mutation in exon 17 of the poly(ADP-ribose) polymerase 1 (*PARP1*) gene (V762A) that was present in all 4 C9-ALS lines, but present in only one of the controls. As this was one of the genes found in the nodes of the integrated network, it is possible that the significant changes observed in the RNA-Seq data may be more likely due to this genomic variant rather than a consequence of the HRE in *C9ORF72*. Further, we have no reason to believe that this variant is a haplotype that is associated with the *C9ORF72* expansion. Therefore, we sought to relate the WGA to the omics results to better determine which genes were differentially expressed due to the HRE in *C9ORF72* and which might be due to line-specific genetic variation at other loci. We focused on exonic variants and found 7,235 nonsynonymous variants that were enriched in either the controls or ALS cases. Then, we compared the genes in which these variants were found to the genes that were found to be differentially expressed (FDR < 0.1, which corresponds to p-value < 0.015) in C9-ALS or control samples by RNA-Seq. We observed 801 variants (including missense, stop gain, start loss, splicing, frameshift) in genes that were differentially expressed. To examine if these subset of differentially expressed genes were significantly correlated to the presence of the variant, we performed linear regression. After voom normalization of the gene expression counts, using the limma package, a linear model was fit to each normalized gene expression-variant comparison. Adjusted R^2 and Benjamini-Hochberg adjusted p-values were calculated for each linear fit. This linear regression analysis revealed 69 variants that could be influencing the expression of 56 genes and confounding the identification of *C9ORF72* ALS-specific gene expression differences. Seven of these genes were found in the final network analysis, but some discordance can be seen in

the genotype-expression comparisons, which could be due to the limited number of samples for the regression analysis. To further try and determine whether genetic variants in our samples were confounding the identification of an ALS signature, we compared the variants that were enriched in either the controls or cases to known brain-specific eQTLs from the xQTL database (Ng et al., 2017). There were 73,142 variants in our samples that overlapped with significant known brain eQTLs that represented 5,292 genes; of these genes, 114 overlap genes were found to be significantly differentially expressed in the ALS versus control cases. 19 of these variants were found in all cases of one group only versus the other group, e.g. all ALS cases and no controls or no ALS case and all controls. These 19 variants are known eQTLs for 7 genes that were also found in our RNA-Seq analysis to be differentially expressed between ALS and control groups; one of which, integrin subunit alpha V (*ITGAV*), was identified as dysregulated in each primary assay, WGA, network and as a fly modifier gene. These analyses demonstrate that the known brain eQTLs are likely to have at most a modest effect on the expression differences between *C9ORF72* and control lines in our study.

B.3.9 An "omics integrator" reveals novel C9-specific pathogenic pathways

We next investigated potential functional links between the data, using a strategy implemented in Omics Integrator (Tuncbag et al., 2016). This approach begins by jointly analyzing the epigenomic and gene expression data to identify transcriptional regulators, which tend to be difficult to detect using mass-spectrometry. Omics Integrator then uses network optimization to search a vast database of protein-protein interactions to discovery, *de novo*, pathways linking the experimentally determined proteomic data and the inferred transcription factors.

B.3.10 Identification of transcriptional regulators

Potential transcriptional regulators were identified using *de novo* DNA motif analysis. To capture regulators mediating changes in chromatin accessibility, we searched for motifs that are enriched in differentially accessible peaks. We also searched within peaks that changed in accessibility and were near differentially expressed genes to identify transcriptional regulators that drive changes in gene

expression. Peaks that were less accessible in C9-ALS samples were enriched for several TFs including Nuclear Factor I (NF1) family that controls the onset of gliogenesis in the developing spinal cord and LIM Homeobox (LHX) TFs that regulate expression of axon guidance receptors (Figure B-4A) (Deneen et al., 2006; Palmesino et al., 2010). Conversely, peaks that were more accessible in C9-ALS samples were enriched for AP-1, RUNX2, and TEAD4. Altered AP-1 activity, which was independently predicted by IPA of the transcriptomics data, has previously been described in *SOD1* mouse models (Bhinge et al., 2017). Notably, we found that RNA transcripts corresponding to motifs enriched in C9-ALS peaks are upregulated in C9-ALS samples, while transcripts corresponding to motifs enriched in control peaks are downregulated in C9-ALS samples (Figure B-4B). These results suggest that epigenetic changes could be driven by differences in expression of transcription factor transcripts.

B.3.11 A network of *C9ORF72*-induced changes

In the next phase of the integration, we combined the transcriptional regulators inferred from RNA-Seq and ATAC-seq with the proteins detected in mass spectrometry. Our approach sought to discover, *de novo*, the cellular pathways that are differentially active between C9 and control lines. The challenge is to go beyond the limited information available in annotated pathways while still avoiding an uninterpretable network containing thousands of interactions. Our approach searches for previously reported protein-protein interactions that connect, directly or indirectly, our proteomics and transcriptional regulatory data. The method considers the strength of experimental evidence supporting each reported protein-protein interaction from the database and the strength of evidence supporting our own data.

Omics Integrator was used to search for connections among 376 predicted TFs and differentially expressed proteins. After optimization and filtering for robustness, the network retained 291 of these proteins and added 83 other proteins that were closely connected by physical interactions. The resulting 374 node network is shown in Figure B-5A, with nodes organized by cellular compartment.

To evaluate the performance of our algorithm, we assessed the network for enrichment of genes previously associated with ALS. We found strong enrichment for

ALS-associated proteins (Figure B-5A bolded borders; p-value = 4.0E-13). We also found that the 83 proteins added by Omics Integrator were also enriched for ALS associated genes (p-value = 2.4E-3), providing confidence that our method is predicting disease-relevant proteins and pathways.

In order to understand the function of the identified network we scored it using categories from Gene Ontology. Enrichment analysis revealed significant dysregulation of ECM, in line with our transcriptomic, proteomic, and epigenomic results. Furthermore, our network was enriched for proteins belonging to cytoskeletal organization and RNA metabolism pathways (Figure B-5A,B), both previously implicated in ALS. For instance, the nuclear-cytoskeletal compartment contains cofilin (CFL1), a known interaction partner of C9ORF72 that modulates actin dynamics in motor neurons (Sivadasan et al., 2016). LIMK1, a kinase that phosphorylates CFL1 also appears in the network and is known to also phosphorylate MMP14 (found in the cytoskeletal-plasma membrane compartment in Figure B-5A,C), an endopeptidase that degrades ECM components (Lagoutte et al., 2016). Proteins involved in microtubule organization (PPP2CA, MAP1B, tubulin) are also represented in the cytoskeletal component of the network. PPP2CA, a major phosphatase for microtubule-associated proteins and a known binding partner of C9ORF72, has been shown to activate MAP1B which in turn tyrosinates tubulin (Coyne et al., 2014). Our network also features mitochondrial proteins that are involved in responses to oxidative stress. Mutations in *PARK7* have been linked to ALS, and its knockdown has been shown to increase disease severity in *SOD1* mouse models (Hanagasi et al., 2016; Lev et al., 2015). Furthermore PINK1, a PARK7 mitochondrial cofactor, plays a role in axonal transport of mitochondria (Moller et al., 2017). Lysosomal dysfunction has also been implicated in ALS (Hardiman et al., 2017). Small GTPase RAB39B plays an important role in the initiation of autophagy via C9ORG72's GDP-GTP exchange factor activity (Corbier and Sellier, 2017). UBQLN4, linked to ALS and found in the cytoplasmic component of the network, may assist in maturation of autophagosomes (Edens et al., 2017).

The network also reveals potentially pathological interactions between differential proteins and predicted transcriptional regulators. SUMOylation via SUMO2 is a post-translational modification process that can affect structure, localization, activity, and

stability of substrates. Specifically, SUMOylation of POU5F1 (Oct4) and PAX7 enhances their stability and transactivity, while SUMOylation of JUN (AP1 family), ETS1, and RUNX2 reduces their stability and transactivity (Bossis et al., 2005; Ji et al., 2007; Luan et al., 2013; Wei et al., 2007). Notably, SUMO2 protein is downregulated in ALS samples, and the activity of these transcriptional regulators following SUMOylation is concordant with their predicted activity. SUMOylation's role in affecting the stability of hnRNPs and localization of actin components to the nucleus has previously been reported (Hofmann et al., 2009; Lee et al., 2012). Our analysis provides evidence that SUMOylation may have substantial influence on transcriptional regulation in C9-ALS motor neurons.

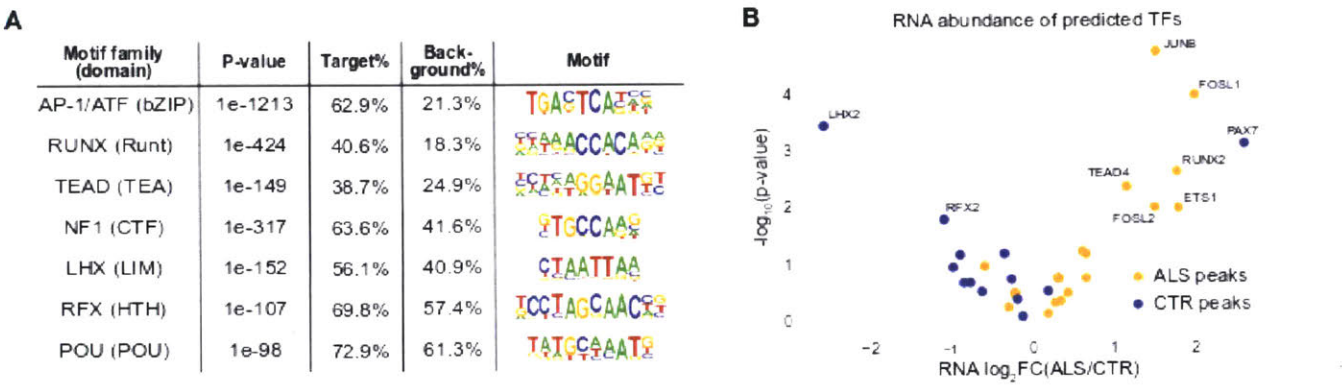

Figure B-4 (A) Transcription factor families that are predicted to be differentially active between ALS and control samples. Orange motifs are predicted to be more active in ALS and blue motifs are predicted to be more active in controls. (B) A volcano plot of RNA abundance for each predicted TF shows that TFs that are predicted to be active in ALS are also more highly expressed in ALS samples, while TFs that are predicted to be active in controls are less expressed in ALS samples.

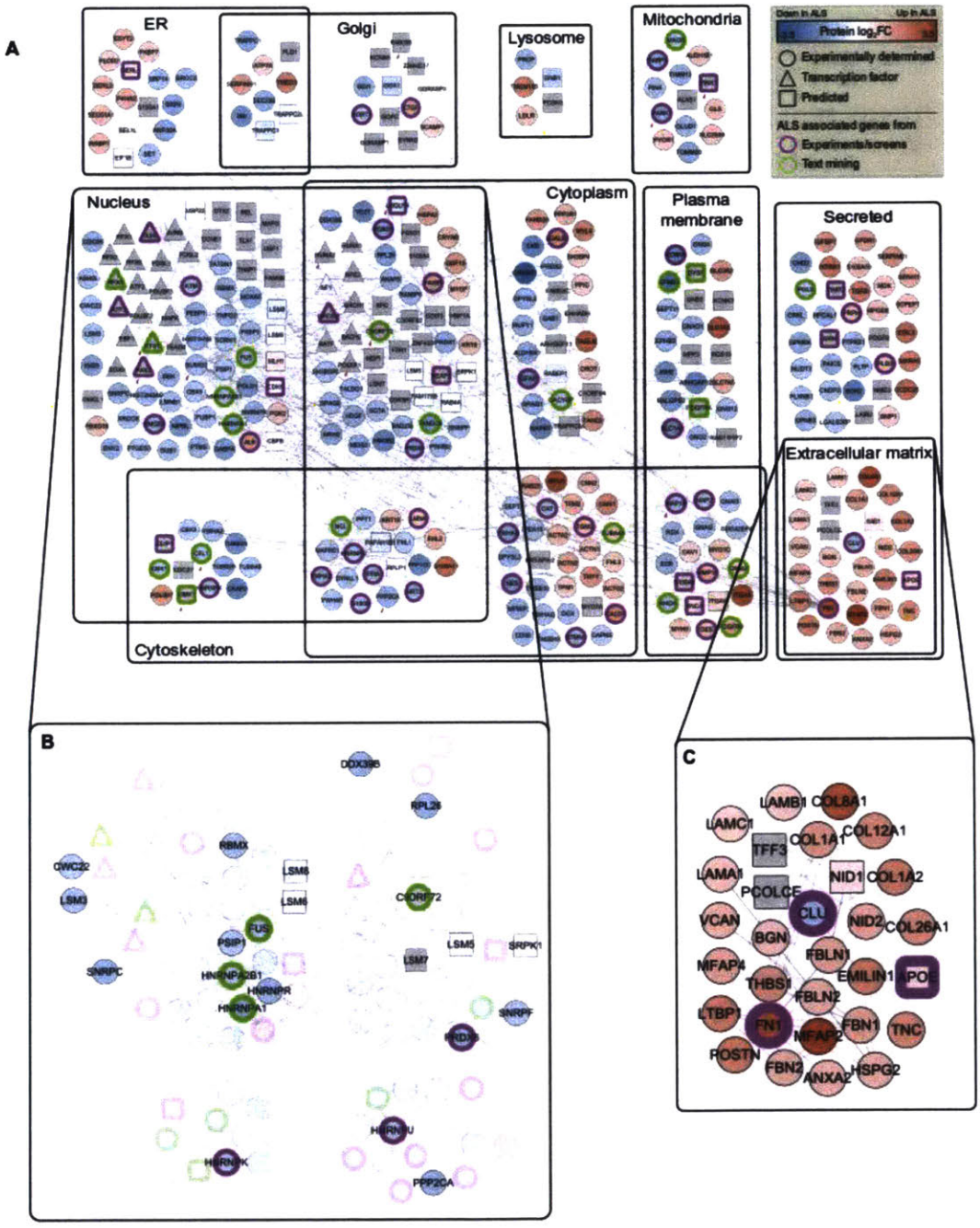

Figure B-5 (A) Integrative analysis reveals a network of 374 proteins organized by subcellular location, of which 264 are experimentally determined from proteomics (circles), 27 are predicted transcription factors, and 83 are other proteins that were closely connected by physical interactions. Borders indicate ALS-associated genes from experiments or screens (purple) and text mining (green). (B) A zoomed in view of the nucleus compartment displaying genes with RNA metabolism functions. (C) A zoomed in view of the extracellular matrix compartment.

B.3.12 Validation of key pathways from the literature and using a fly screen

Many of the pathways identified using the Omics Integrator could also be found by searching the current ALS literature as described above. In addition, there were novel pathways that had not previously been reported as disrupted in C9-ALS including the extracellular matrix and cytoskeleton. In order to validate our "integrated omics" list generated from control and C9-ALS iMNs in vivo, we conducted an RNAi-based screen in a Drosophila model of G4C2-mediated neurodegeneration (Xu et al., 2013). In this model, over-expression of 30 G4C2 repeats in the eye leads to age-dependent photoreceptor neurodegeneration, and genetic pathways identified as modifiers of fly eye degeneration have proven to be relevant to *C9ORF72*-associated neurodegeneration in mouse and human iPSC-derived neuron models (Xu et al., 2013; Zhang et al., 2015). A total of 293 fly genes corresponding to 284 human genes were knocked down in the G4C2 fly model and their ability to modify (suppress or enhance) the rough eye phenotype was scored (Figure B-5B). When available, multiple RNAi lines were tested. Of those, about 20% enhanced and 15% suppressed C9 toxicity with a score of at least +/-1 respectively. The remainder showed little or no effect on eye degeneration and approximately 2% resulted in lethality. There was no particular relationship between the proteomic changes in iMNs and the phenotypic effect of knocking-down the gene in the fly. The results from the fly screen confirm that a subset of genes/proteins, identified through our integrated omics approach, are modifiers of *C9ORF72* G4C2-repeat-mediated toxicity. Furthermore, the altered expression of modifier genes/proteins is likely to contribute, at least in part, to *C9ORF72*-mediated toxicity.

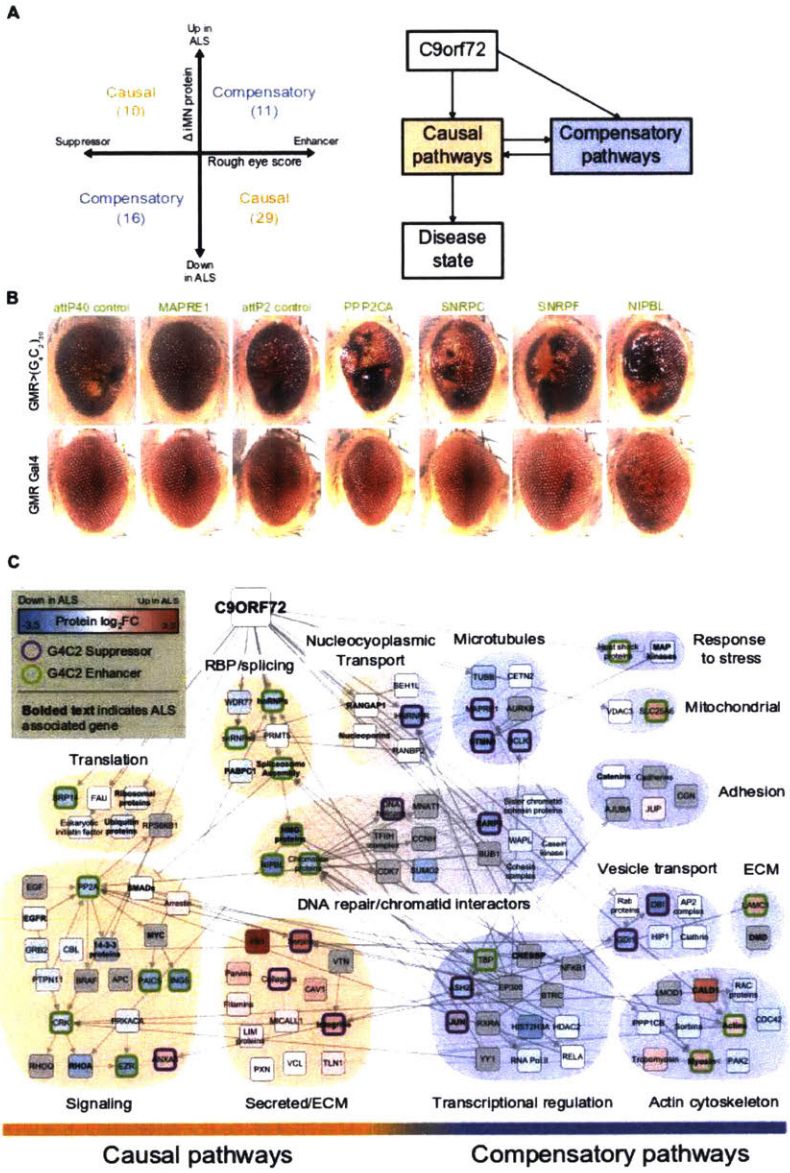

Figure B-6 (A) Left: Each gene that was tested in the fly model are sorted into causal or compensatory categories using its fly phenotype and change in protein values in iMNs. Right: A schematic showing the interplay between causal and compensatory pathways that eventually result in the disease. (B) The effect of genetic manipulations on external eye morphology and depigmentation in G4C2 expressing flies. (C) Causal and compensatory genes from A were connected via intermediate genes and the resulting network was organized by cellular process. Proteins from the same families were consolidated into a single node for readability. The borders indicate whether the gene is a G4C2 suppressor (purple) or enhancer (green). Bolded names indicate ALS-associated genes.

B.3.13 Characterization of putatively causal and compensatory pathways

We leveraged the fly results to explore the potential causal roles of proteins that changed in the iMN data. Based on omic data alone, where specific genes, proteins and pathways are identified as up- or down-regulated, it is not possible to determine whether a difference in ALS versus control motor neurons is part of the toxic effects of the *C9ORF72* expansion or whether it represents a compensatory process. However, we can begin to resolve this ambiguity using the results of the RNAi screens carried out in the fly model of the repeat expansion above. For example, in the simplest case, if a protein is upregulated in C9-ALS motor neuron cultures and knocking it down suppresses eye degeneration in the fly, the ALS-induced change(s) were likely deleterious. We refer to such C9-induced changes as "causal." By contrast, if knock-down of the same protein resulted in enhanced eye degeneration, the ALS-induced change(s) are more likely to be part of a compensatory adaptation. In total, we found 39 causal and 27 compensatory genes (Figure B-6A,B).

We developed an integrative approach to discover the functional interactions among these genes and their underlying roles in ALS pathology. Specifically, we built networks connecting these proteins using directed interactions gathered from two public pathway databases, KEGG and Reactome, and grouped the resulting proteins by functional categories (Figure B-6A).

This approach revealed several causal pathways (Figure B-6C) that were previously known to be dysregulated by the mutated form of *C9ORF72*, such as RNA splicing and nuclear transport (Robberecht and Philips, 2013; Zhang et al., 2015). The altered proteins in these pathways include ALS and associated genes such as hnRNPA1, FUS (located in the Spliceosome Assembly node), and RanGAP1. Other pathways emerged as causal that have been less thoroughly examined in the context of *C9ORF72*. These pathways include signaling pathways such as EGF signaling and SMAD signaling (eg: EZR and CRK), with a hub centered around phosphatase PP2A. The approaches used here also highlighted a novel, causal set of ECM-related pathways and genes including integrins, collagens, and serpins. Within these networks based on the fly data, a number of pathways are likely to represent compensatory changes. For instance, the observed increases in the cytoskeletal proteins like actin,

myosin and tropomyosin, increases in heat shock proteins and decreases in RAC proteins and other proteins relating to GTP/GDP exchange are compensatory. Our approach also begins to reveal interactions between different processes. For example, the putatively causal toxic changes in the nucleocytoplasmic transport or oxidative stress are connected to potentially compensatory changes in DNA repair pathways. Finally, regulation of causal and compensatory processes can be elucidated using this approach (Figure B-6C). For instance, while ECM/secreted proteins fall into causal pathways, cell adhesion protein changes are largely compensatory, as is dysregulation of Laminin C1, which is a component of the basal lamina and is secreted and incorporated into ECM matrices as an integral part of the structural scaffolding in tissues.

B.4 Discussion

iPSC models finally offer a way to map the initiation and execution of pathology in specific diseases of the central nervous system (CNS). This is clearly required given the lack of neurologically active drugs despite years of investment from both industry and academia. Many groups have now been able to generate iPSCs from patients with neurological disease-causing mutations and have shown specific phenotypes in the dish (Huntington's Disease consortium, Parkinson's Disease genetic cases) and there have been two recent studies showing a stress-induced phenotype in C9 iPSC-derived motor neurons and an overall cell death and reduced fiber outgrowth phenotype in a range of ALS cases not including *C9ORF72* (Fujimori et al., 2018; Shi et al., 2018). In another report, increased activity in motor neurons from ALS patients in the dish led to a drug trial with retigabine that is currently still underway (McNeish et al., 2015). Interestingly, all of these studies were focused on discovering physical *in vitro* phenotypes such as cell death or reduced fiber outgrowth, which may or may not be relevant to drug intervention in patients. One of the key difficulties in these studies has been an incomplete picture of the earliest and most significant changes that occur during pathogenesis.

With the premise that dysfunction of molecular pathways in specific cell populations in the brain leads to neurodegeneration, we have established a

comprehensive, quantitative molecular phenotyping approach using a human iPSC technology platform to study molecular signatures of CNS cell types focusing on iPSCs from patients with *C9ORF72*, given its prevalence as a genetic cause of ALS and its dominant phenotype (Brown and Al-Chalabi, 2017). We have used genomics, transcriptomics, epigenomics, high-content quantitative proteomics, and single-cell imaging technologies to characterize human motor neuron cultures from *C9ORF72* ALS patients, under strict quality control including the use of parallel cultures for each assay, metadata standards, and analytical pipelines. A computational pipeline was used to integrate the diverse molecular data sets and identify the most significant regulated pathways in patient cells. This "Omics Integrator" software uses network approaches to integrate diverse data types into coherent biological pathways that can avoid some of the pitfalls associated with analyzing single data types and uncover novel pathways that are not annotated in existing databases (Tuncbag et al., 2016). This approach is validated by the strong statistical enrichment and the comprehensive number of hits it recovered that are consistent with published literature for *C9ORF72* ALS. At the same time, the approach revealed functional links among the disparate data, including identifying many transcriptional regulators.

A challenge in using multi-omic data sets is understanding how the direction of a change impacts disease pathogenesis. This is perhaps one of the greatest difficulties – e.g. understanding if the observed changes are conducive to the course of the disease or a cellular attempt at a homeostatic response to physiological insults. Using Drosophila genetics guided by the outcome of the integrated networks, it has been possible to not only validate the specific genes and proteins involved, but also to discern probable effect and whether altered expression or activity would be predicted to promote disease pathogenesis or serve as a compensatory response. The results of these studies provide a unique data source and methods that can be utilized in the study of ALS and other neurodegenerative diseases.

Our analysis reveals a complex system of interweaving relationships among causal and compensatory pathways. In some cases, such as the ECM, causal and compensatory roles were found to exist even within the same pathway. Though the literature on the ECM's role in neuronal function and disease progression is limited,

several studies have described neuroprotective properties of the ECM (Suttkus et al., 2016). Our analysis suggests that, while ECM components are broadly upregulated in ALS, individual components of the ECM may have very different downstream consequences. For example, knocking down some genes like *LAMC1* and *DMD* enhances toxicity in fly eyes while knocking down other ECM components like serpins, collagens and integrins suppresses toxicity. One mechanism through which extracellular signals within the ECM may be internalized is through integrin signaling. Integrin activation mediates molecular coupling of CAS and Crk, and the resulting complex has been shown to regulate the actin cytoskeleton (Chodniewicz and Klemke, 2004). Interestingly, integrins and CRK were both found to be pathogenic, while actin cytoskeletal components were compensatory, which suggests ECM pathogenicity is transmitted via some non-cytoskeletal pathway.

It is also important to recognize that the classification of changes as "causal" or "compensatory" is far from definitive. Not all results from the fly necessarily translate to human cells and tissue. Furthermore, our simple binary classification does not capture complicated situations in which there may be non-linear effects of gene expression on phenotypes. However, these first attempts at relating many different aspects of cell functioning are the starting blocks for further studies and enable for the first time the development of a holistic view of cell functioning in the face of a pathogenic repeat that causes ALS.

This integrative approach is well suited for the task of hypothesis generation. For instance, our results suggest that DNA repair pathways are a compensatory response to either nucleocytoplasmic transport or oxidative stress. In addition to providing insight into how these pathways interact, our analysis also identifies proteins that are attractive targets such as MAPRE1. While we acknowledge there are some limitations of integrating data across human *in vitro* and fly *in vivo* models, this approach provides a much-needed basis for establishing causality and generating testable hypotheses.

An additional benefit in having transcriptomic and proteomic data together with WGS is the ability to integrate these data sets and identify whether a given DNA sequence change causes altered expression of the gene or altered levels of the protein. Using the data set here, we have integrated WGS with RNA-Seq data to begin to

evaluate eQTLs that may be meaningful to disease as a causal modifier versus altering gene expression as a consequence of ALS. Future studies will expand this analysis across each of assays and extend to larger data sets from additional ALS subjects.

Based on the hypotheses generated through the generation of integrated networks and potential causality suggested by the fly data, next steps will include testing whether modulation of these pathways in the iPSC neurons can impact key pathogenic features of *C9ORF72* such as the nuclear pore deficit and formation of dipeptide repeats. Using robotic imaging, there is now the potential to use reporters to query specific networks or processes (e.g. ECM) in future studies (Finkbeiner et al., 2015). Finally, validation in human brain tissue can provide insights as to the relevance of specific pathways identified here to represent very early changes to later stage disease pathology. The current study gathered a wide range of critical information, but was underpowered with regard to numbers of patients and made no connection to the complex clinical course of the disease. Currently we are producing 1000 iPSC lines from patients with all types of ALS (including *C9ORF72* mutation carriers) and performing a similar analysis. In addition, the clinical history of each patient will be combined with the Omics Integrator to give more resolution on how molecular changes may impact the clinical course of the disease. However, the core techniques and integrated approach of the current report along with the first set of data suggesting a molecular signature for C9 ALS provide a strong framework for this new "big data" approach to learning more about the causes and treatments of diseases such as ALS.

B.5 Methods

Generation and Characterization of iPSC Lines

The 3 control lines (termed 25iCTR, 83iCTR, 00iCTR) and 4 iPSC lines (termed 29iALS, 52iALS, 30iALS, 28iALS) were generated using episomal plasmids and characterized as previously described (Sareen et al., 2013). Human control fibroblast cell lines were obtained from the Coriell Institute for Medical Research. The Coriell Cell Repository maintains the consent and privacy of the donor of fibroblast samples. Fibroblasts from C9ORF72 ALS patients (28iALS-n2, 29iALS-n1, 30-iALS-n1, and 52iALS-n6) were derived at Washington University of St. Louis. Healthy control

fibroblasts (00iCTR: GM05400; 83iCTR: GM02183) were obtained from the Coriell Institute for Medical Research. All the cell lines and protocols in the present study were carried out in accordance with the guidelines approved by institutional review boards at the Cedars-Sinai Medical Center and Washington University at St. Louis. Studies were performed under the auspices of the Cedars-Sinai Medical Center Institutional Review Board (IRB) approved protocol Pro00028662 and Pro00028515. The reprogramming and characterization of iPSC cell lines and differentiation protocols in the present study were carried out in accordance with the guidelines approved by Stem Cell Research Oversight committee (SCRO) and IRB, under the auspices of IRB-SCRO Protocols Pro00032834 (iPSC Core Repository and Stem Cell Program), Pro00024839 (Using iPS cells to develop novel tools for the treatment of SMA) and Pro00027006 (Cell and Tissue Analysis for Neurologic Diseases; Robert Baloh). Appropriate informed consents were obtained from all the donors. To protect donor privacy and confidentiality, all samples were coded and de-identified in this study.

Extensive quality control processes were implemented, including testing iPSC precursors and final neuronal samples (motor neuron cultures) for purity and their identity by short-tandem repeat (STR) analysis performed by a third-party company before the samples were distributed. G-band karyotyping was performed to ensure that iPSCs maintained normal karyotypes. The parental tissue (fibroblasts), reprogrammed iPSCs and the differentiated iMNs prior to performing assays were submitted to IDEXX BioResearch for DNA fingerprinting and STR analysis to confirm donor identity. Each of the iPSC lines used in this study had unique genetic profiles and the genic profiles of the samples and their source tissues were identical. Additionally, the test confirmed the samples to be of human origin and detected no mammalian interspecies contamination. The Cedars-Sinai iPSC Core Facility created a working cell bank of iPSC-derived motor neuron precursor spheres (iMPS) for C9-ALS and control subjects.

Whole Genome Sequencing and Analysis

DNA was extracted from iPSC lines made in the laboratory of Clive Svendsen and Dhruv Sareen using the QIAamp DNA Blood mini Kit (Qiagen; 51104) as per the manufacturer's instructions. A minimum of 1µg of unamplified, high molecular weight,

RNase treated DNA with absorbance values of OD260/280 1.7- 2.0 and OD260/230 > 2.0, was sent to The New York Genome Center for sequencing on the Illumnia X10. Sequence data was processed on NYGC automated pipeline. Paired-end 150 bp reads were aligned to the GRCh37 human reference using the Burrows-Wheeler Aligner (BWA-MEMv0.7.8) and processed using the GATK best-practices workflow that includes marking of duplicate reads by the use of Picard tools (v1.83, http://picard.sourceforge.net), local realignment around indels, and base quality score recalibration (BQSR) via Genome Analysis Toolkit (GATK v3.4.0) (New York Genome Center) (DePristo et al., 2011; McKenna et al., 2010).

The variant calls from NYGC were assessed by examining the actual reads for alignment issues and spot-checking the BAM files for specific variants in IGV and assessed they were of good quality. The VCFs were converted into GVCFs and performed custom annotation and intersected a subset of the omics data (RNA-Seq, ATAC Cluster) with the WGS data.

The annotation pipeline was customized to incorporated elements from ANNOVAR and KGGseq from which a report was generated, including genotypes for all samples (Li et al., 2012; Wang et al., 2010). These reports are available upon request. The following annotation was used: For genes and exonic variants that have clinical significance, we incorporated the Clinical Genomic Database (CGD), the Online Mendelian Inheritance in Man (OMIM), ClinVar, and genes listed in the American College of Medical Genetics and Genomics (ACMG) as well (Amberger et al., 2015; Green et al., 2013; Landrum et al., 2016; Solomon et al., 2013). Intervar, which is based upon the ACMG and AMP standards and guidelines for interpretation of variants was also incorporated. This tool uses 18 criteria to prescribe the clinical significance and classifies based on a five- tiered system (Farrer et al., 1997). To flag ALS genes, we incorporated ALS gene lists and variants from ASLoD (http://alsod.iop.kcl.ac.uk/), a highly curated list from Dr. John Landers and ALS associations from the DisGeNet database (Abel et al., 2013; Piñero et al., 2017). We also incorporated functional prediction by using in silico prediction from nine programs, including the databases, such as SIFT, PolyPhen2, and MutationTaster and as in Li et al., 2013 for each variant (Chun and Fay, 2009; Li et al., 2013; Schwarz et al., 2010; Sim et al., 2012). As well,

additional databases were included that assess the variant tolerance of each gene using the RVIS and the Gene Damage Index (GDI) and are adding LoFTool (Fadista et al., 2017; Itan et al., 2015; Petrovski et al., 2013). Gene expression: For variants in genes that are highly expressed in the brain, we provided these data from the Human Protein Atlas (http://www.proteinatlas.org) and expression data from GTex portal (2013, 2015; https://gtexportal.org/home/) for the cortex and spinal cord (Uhlén et al., 2015). Frequency information from three databases on all known variants from ExAC, the NHLBI Exome Sequencing Project (ESP), and the 1000 Genomes Project (Auton et al., 2015; Lek et al., 2016; Tennessen et al., 2012).

A separate annotation pipeline was developed for variants that are in intergenic and regulatory regions. We report the variant as found next to the closest gene, these are either intronic, upstream and downstream (up to 4 KBs from the start and stop of a gene) and 5' and 3' UTRs. The annotation used came from: RegulomeDB which annotates variants with known or predicted regulatory elements such as transcription factor binding sites (TFBS), eQTLs, validated functional SNPs and DNase sensitivity (Boyle et al., 2012). The source data comes from ENCODE (2004; 2012) and GEO (Barrett et al., 2009). We also included other regulatory databases such as Target Scan is an algorithm that uses 14 features to predict and identify microRNA target sites within mRNAs and miRBase (Agarwal et al., 2015; Griffiths-Jones, 2004, 2005; Griffiths-Jones et al., 2008).

Differentiation of iPSCs into Motor Neurons

Control and ALS iPSCs were differentiated into motor neurons based on a combination of previous models established for rapid neural differentiation (Figure B-1A) (Sances et al., 2016). Briefly, iPSCs were grown to near confluence devoid of spontaneous differentiation under normal maintenance conditions prior to the start of differentiation. Neuroectoderm specification of iPSCs was induced by removal of mTeSR1 media and addition of defined neural differentiation media (NDM) +LS composed of IMDM supplemented with B27 + vitamin A (2%), N2 (1%), Non-Essential Amino Acids (NEAA, 1%) and penicillin-streptomycin-amphotericin (PSA, 1%) along with LDN193189 and SB431542 [LS] - as a combination of small molecule inhibitors of

SMAD pathway, BMP type 1 receptors (ALK2/3) TGF-beta superfamily type 1 activin receptor-like kinase (ALK) receptors (ALK4/5/7)]. Colonies were dissociated into single cells with Accutase and uniform aggregates were formed in sterilized V-bottom 384-well PCR plates with 20,000 cells/well. Uniform neural aggregates were formed by seeding in NDM+LS in presence of Matrigel and centrifuging for 5 minutes at 200g. The aggregates were maintained in this media for 5 days. The culture medium was replenished every 2 days. On day 5, the aggregates were gently isolated from the plates using Accutase, and the uniform sized neural aggregates were then plated on laminin-coated 6-well plates. After 7 days (day 12), media were changed to a motor neuron specification medium (MNSM) generating caudo-ventralized MN precursors by addition of all-trans retinoic acid (ATRA) and the sonic hedgehog agonist, purmorphamine (PMN), brain-derived neurotrophic factor (BDNF), glial cell line-derived neurotrophic factor (GDNF), ascorbic acid (AA) and dibutyryl cyclic adenosine monophosphate (db-cAMP). Over the next 4 to 8 days, neural rosettes formed and were lifted at day 16 to 20 and subsequently cultured in suspension low-attachment flasks for a further 8 days. Selected rosettes were switched to a motor neuron precursor expansion media (MNPEM) containing ATRA, PMN, and the mitogens epidermal growth factor (EGF) and fibroblast growth factor (FGF2). After an initial 8 days in the expansion medium the generated induced motor neuron precursor spheres (iMPS) were further expanded by weekly chopping for 5 weeks (passages) and cryopreserved prior to initiation of terminal differentiation stage. These iMPS were cryopreserved into lots for later generation of iMPS-derived motor neurons (iMNs) for omic analysis or to send to imaging centers for cell death assays and live-cell imaging. In order to induce terminal motor neuron differentiation, the iMPS were fully dissociated with Accutate and seeded on laminin-coated 6-well plates, and matured in Stage 1 motor neuron maturation medium (MNMM Stage 1) consisting of NDM supplemented with ATRA (0.1µM), PMN (1µM), db-cAMP (1µM), ascorbic acid (AA; 200ng/ml), Notch signaling γ-Secretase Inhibitor, DAPT (2.5µM), BDNF; 10ng/ml and GDNF; 10ng/ml for 7 days. Then cultures were switched to maturation medium stage 2 (MNMM Stage 2) containing Neurobasal, 1% NEAA, 1% N2, 0.5% GlutaMax, db-cAMP (1µM), ascorbic acid (AA; 200ng/ml), BDNF; 10ng/ml and GDNF; 10ng/ml for another 14 days. Mature iMN cultures were harvested and screened

at 21-days post plating. These conditions allowed for motor neuron differentiation under serum-free conditions. All differentiating cultures were maintained in humidified incubators at 37°C (5% CO_2 in air).

Immunocytochemistry

Human iPSC-derived motor neuron cultures were plated on optical-bottom 96-well plates (Thermo, #165305) and subsequently fixed in 4% paraformaldehyde for 15 minutes. Cells were blocked in 5% normal donkey serum with 0.1% Triton X-100 in phosphate buffered saline (PBS) and incubated with primary antibodies either for either 1 hour at room temperature or overnight at 4°C. Cells were then rinsed and incubated in species-specific AF488, AF594, or AF647-conjugated secondary antibodies followed by Hoechst 33258 (0.5µg/mL; Sigma) to counterstain nuclei. Cells were imaged using Molecular Devices ImageExpress Micro high-content imaging system or using Leica microscopes (Fuller, Mandefro et al. 2015). Primary antibodies used were as follows: mouse anti-SMI32 (Covance, 1:1,000); mouse anti-TuJ1 (β3-tubulin) (Sigma; 1:1,000-1:2,000); rabbit anti-glial fibrillary acidic protein (GFAP, Dako; 1:1,000); mouse anti-Map2a/b (Sigma; 1:1,000); rabbit anti-nestin (Millipore; 1:2,000).

Longitudinal Single Cell Analysis

To generate iMNs for automated robotic imaging, frozen vials of iMPS were obtained from Cedars-Sinai. iMPS were quickly thawed at 37°C and then dissociated into single cells with Accutase at room temperature. To ensure single-cell dissociation, Accutase-treated iMPS were gently pipetted up and down and passed through a 30mm Cell Strainer and washed with PBS. The Accutase was removed with centrifugation (200g for 5 minutes at room temperature) and cells were gently resuspended in MNMM Stage 1 media [NDM supplemented with 0.1mm all-trans RA (ATRA; Sigma), 1mM PMN; Stemgent, 10ng/ml of BDNF; R&D Systems, 10ng/ml GDNF; R&D systems, 200ng/ml AA, 1mM db-cAMP, and 1% Antibiotic-Antimycotic; LifeTech. Matrigel was added to the cell suspension (1:100 by volume) and cells were seeded on a Matrigel-coated 96-well plate at 50,000 cells/well. After 4 hours, the media was replaced. The following day, cells were fed with MNMM Stage 1 media supplemented with 2.5mM

DAPT (Tocris) every other day. On Day 8, cells were switched into MNMM Stage 2 media [Neurobasal media supplemented by 1% NEAA, 0.5% Glutamax, 1% N2, 10 ng/ml BDNF, 10 ng/ml GDNF, 200ng/ml AA, 1mM db-cAMP, and low-dose Cytarabine (AraC) at 0.1mM (used to block residual glial cell proliferation) and fed every other day. On Day 20, cells were transfected with 500ng/well of motor neuron morphology marker HB9-green fluorescent protein (GFP) plasmid using Lipofectamine 3000 (Thermo Fisher Scientific) according to manufacturer's instructions. A half media change was performed every other day until the cell culture plate was imaged and fixed. At Day 25, the cells were imaged every 12 hours for 7 days.

In a subset of experiments, nucleofection was used to introduce plasmids into iMNs during suspension stage. In the beginning, iMPS were maintained in T75 flasks and allowed to expand in MNPEM media followed by MNMM Stage 1 media change at Day 0. The iMPS spheres were then subjected to MNMM Stage 2 media at Day 7 and transfected on Day 12 using Human Stem Cell Nucleofector Kit 2 (Lonza Amaxa; Cat # VPH-5022) per manufacturer's guidelines. Briefly, iMPS spheres were washed 4 times in PBS solution and suspended in 100µl of nucleofector solution mixed with 5µg of HB9-GFP and human synapsin promoter driven Syn-mApple plasmid. The suspension mix was transferred to a provided cuvette from the kit and nucleofected with a Nucleofector apparatus (Amaxa). Cells were transfected using A-033 pulsing parameter and were immediately transferred into the original T75 flasks containing MNMM Stage 2 media. The iMPS spheres were allowed to recover for two days and were then dissociated into single cells with Accutase. A similar cell dissociation procedure was performed as mentioned previously and the cells were plated at 75,000 cells/well in 96-well plate. A half media change was performed every other day until the cell culture was imaged and fixed. At Day 25, the cells were imaged every 12 hours for 7 days.

Plates of transfected cells were maintained at 37°C and 5% CO2 in a robotic incubator until imaged. Within a 37°C environmental chamber, a robotic arm transferred each plate to the stage of the microscope for automated image acquisition. The protocol uses a fiduciary mark on the plate for alignment, performs automated focusing, and then collects a series of fluorescence images of adjacent fields from each well. Images from a single well were stitched together into montages. A custom cell identification algorithm

generates a single cell mask for each montage. The program then aligns montages from the same well at sequential time points and assigns unique numbers to individual cells that are tracked during the experiment. The survival time for each cell of each well is determined and quantified from the images using methods described previously (Skibinski et al., 2014). Kaplan-Meier curves are constructed from individual survival times of cells from each well and survival of cohorts of cells from each well is compared to each other with survival or time-to-event analysis.

RNA-Seq

Total RNA was isolated from each sample using the Qiagen RNeasy mini kit. RNA samples for each subject (control or disease) were entered into an electronic tracking system and processed at the University of California, Irvine GHTF. RNA QC was conducted using an Agilent Bioanalyzer and Nanodrop. Our primary QC metric for RNA quality is based on RIN values (RNA Integrity Number) ranging from 0-10, 10 being the highest quality RNA. Additionally, we collected QC data on total RNA concentration and 260/280 and 260/230 ratios to evaluate any potential contamination. Only samples with RIN > 8 were used for library prep and sequencing. Library prep processing was initiated with total RNA of 1ug using a Ribo-Zero Gold rRNA depletion and Truseq Stranded total RNA kit. Additionally, ERCC exFold spiked-in controls were used for further QC and downstream data analysis. Briefly, RNA was chemically fragmented and subjected to reverse transcription, end repair, phosphorylation, A-tailing, ligation of barcoded sequencing adapters, and enrichment of adapter-ligated cDNAs. RNA-Seq libraries were titrated by qPCR (Kapa), normalized according to size (Agilent Bioanalyzer 2100 High Sensitivity chip). Each cDNA library was then subjected to Illumina (HiSeq 2500) paired end (PE), 100 cycle sequencing to obtain approximately 50-65M PE reads. After sequencing fastq were subject to QC measures and reads with quality scores (>Q20) collected and analyzed using the pipeline described at http://neurolincs.org/tools/. Briefly, reads were mapped to the GRCh73 reference genome, QCed, and gene expression and differential expression were quantified using the pipeline outlined here: http://galaxy.neurolincs.org/u/terri/p/neurolincs-data-analysis-workflows, using tools HTseq and DESeq2 (Anders et al., 2015; Love et al., 2014).

Normalized and transformed count data were then used for exploratory analysis and DE genes (FDR < 0.1) were used for pathway, network, and gene ontology analysis. These primary data were subject to additional statistical and network-based data analyses using commercial and open-source pathway and network analysis tools, including Ingenuity Pathway Analysis (IPA), GOrilla, Cytoscape, and other tools to identify transcriptional regulators, predict epigenomic changes, and determine potential downstream pathway and cellular functional effects. Significant DEGs (FDR < 0.1) were then analyzed against genes that were found to contain exonic enriched genetic variants from the WGS. The gene expression (voom normalized and transformed values) and genotype variant pairs were analyzed by fitting a linear regression model. Adjusted R^2 and Benjamini-Hochberg adjusted p-values were calculated, significant genes were reported at FDR < 0.1.

Proteomics

Frozen cell pellets were lysed using a combination of lysis buffer containing SDS and sonication. BCA assay was used to determine protein concentration and 125ug of each sample was used in downstream sample processing. Samples were processed following Expedeon FASP protocol. Samples were digested in Trypsin/LysC (Promega) at a ratio of 40:1 to protein concentration at 37°C for 12 hrs. Samples were desalted using MCX micro-elution column (Waters) and samples were dried in speedvac and stored in -20°C until resuspension with Biognosys iRT mixture for acquisition on the SCIEX 6600 over a 45-minute gradient. Samples were acquired in data-dependent acquisition (DDA) mode for library building and in data-independent acquisition (DIA) mode over 100 variable windows similar to acquisition protocols in Kirk et al. and Holewinski et al. (Holewinski et al., 2016; Kirk et al., 2015). DDA files were run through Trans Proteome Pipeline (TPP) using a human canonical FASTA file (Uniprot). A consensus peptide library with decoys was generated. DDA library build principals as described in Parker et al. were utilized to generate a cell specific library, which allowed for more accuracy in matching DIA data to the DDA library during OpenSWATH, as indicated by higher d-scores in PyProphet (Parker et al., 2016). DIA files were mapped onto this library using OpenSWATH and transition level data was compiled with a 1%

FDR cutoff. Downstream summing of transition level data to peptide and protein level data was performed by MAP DIA (Teo et al., 2015). Log2FC data was calculated by MAP DIA and filtered using a 1% FDR, 95% confidence interval and 0.6 abs(log2FC) cutoff to obtain a final list of differentially expressed proteins. For protein quantification, transitions and peptides common to more than one protein were excluded. These data have been further analyzed using commercial and open-source pathway and network analysis tools, including Ingenuity Pathway Analysis and Gorilla, to identify upstream regulators and determine affected cellular pathways.

ATAC-Seq

We used the assay for transposase-accessible chromatin using sequencing (ATAC-Seq) to assess chromatin accessibility and identify functional regulatory sites involved in driving transcriptional changes associated with *C9ORF72*. ATAC-Seq detects open chromatin sites and maps transcription factor binding events in regulatory elements genome-wide, without needing any prior information about which proteins are bound. By correlating ATAC-Seq patterns with other features, such as gene expression, we are able to delineate the fine-scale architecture of the regulatory framework. Chromatin accessibility signatures were generated for each sample individually with detection of differential peaks between disease and control states to generate an initial disease-state signature.

ATAC-seq was carried out as described (Milani et al., 2016). Briefly, cells were lysed in cell lysis buffer (10mM Tris-HCl, pH7.4, 10mM NaCl, 3mM MgCl2, 0.1% IGEPAL CA-630, protease inhibitors) on ice for 5 minutes and centrifuged at 230 rcf for 5 minutes at 4°C. The pellet, containing the nuclei, was re-suspended in 25ul of 1X Tagment DNA Buffer (Illumina). 50K nuclei were subjected to transposase reaction (Nextera - Illumina) followed by DNA purification. The tagmented DNA was PCR amplified using Nextera indexing primers (Illumina) and loaded on 2% agarose gel. Nucleosome-free fragment (175-250 bp) were size selected from the gel and further amplified by PCR to obtain the final libraries. The libraries were sequenced using the Illumina HiSeq 2000 platform (single end, 50 bp). All samples passed quality control checks that included morphological evaluation of nuclei, agarose gel electrophoresis of

libraries, and real-time qPCR to assess the enrichment of open-chromatin sites. The quality of the sequencing was assessed using FastQC and the reads were aligned to GRCh37 genome build using BWA. We identified open chromatin regions separately for each sample using the peak-calling software MACS2 and determined differentially open sites using DESeq2 (FDR < 0.1). Peaks were assigned to unique genes using the default HOMER parameters, and gene ontology analysis was performed using GOrilla (Eden et al., 2009; Heinz et al., 2010; Zhang et al., 2008).

Data Integration

We used a hierarchical strategy for data integration. We inferred transcriptional regulators from the combination of ATAC-seq and RNA-Seq data, and then searched for connections among these transcriptional regulators and those detected directly by the proteomics.

Inferring transcriptional regulators: Accessible chromatin regions, identified by ATAC-seq, were combined with differential gene expression data to predict transcription factors (TFs) that contribute to differences in transcriptomics profiles between C9 and controls. Specifically, we used the union of peaks detected in ALS and control samples to identify peaks proximal (+/-2.5kb) and distal (+/-50kb) to gene transcription start sites (TSS), which were further divided into those with high and low CpG content. A normalized CpG metric was used. We determined the enrichment of known motifs using HOMER. The analysis was performed separately for high and low CpG content peaks near (+/-10kb or 50kb) differentially expressed genes as the foreground and corresponding regions near all known genes as the background.

Network Analysis

We used Omics Integrator to search for previously reported protein-protein interactions that link proteins detected by mass-spectrometry and the inferred transcription factors (Soltis et al., 2017; Tuncbag et al., 2016). Taking a network approach, we represented proteins and TFs as nodes and assigned prizes to them based on their experimental significance. Specifically, protein prizes were assigned according to the fold change between C9 and control samples and prizes for TFs were

assigned according to false discovery rate (see above). We mapped these proteins on a network of physical interactions in which each edge was scored for reliability based on the underlying experimental data. Our algorithm searches for disease-associated subnetworks that retain the maximum prizes while avoiding unreliable interactions which are formalized as the Prize-Collecting Steiner Forest problem. We aim to find a forest solution $F(V_F,E_F)$ that maximizes the objective function:

$$f(F) = \beta \cdot \sum_{v \in V_F} p(v) - \sum_{e \in E_F} c'(e) + \omega \cdot \kappa$$

The first term is the sum of prizes included in *F*, scaled by a model parameter β. The second term is a cost function which serves the purpose of only including a node in *F* if the objective function is minimized. The last term allows for the inclusion of κ trees by introducing a root node v_0 that is connected to every other node with a weight ω. This method not only performs feature selection by filtering out protein prizes that are expensive to connect, but also identifies "Steiner" proteins that were not detected as changing in the experiments, but are strongly implicated by the structure of the interactome. A Steiner node is typically included when its interaction neighbors are significant proteins identified from biological experiments. To avoid a bias toward proteins that have many known interactions (high-degree nodes), we impose a regularization term on edges such that the cost of an edge between nodes a and b monotonically increases with d_a and d_b, the node degrees of a and b:

$$c'(e) = c(e) + \alpha \frac{d_a \cdot d_b}{(N - d_a - 1)(N - d_b - 1) + d_a \cdot d_b}.$$

This regularization term corresponds to the probability that an edge exists between a and b given the number of nodes in the interactome, N, and the degrees of a and b. c(e) is the cost of the edge which is inversely related to the amount of experimental evidence supporting the physical interaction between a and b given by iRefIndex (Razick et al., 2008). Finally, we acknowledge that the algorithm is susceptible to noise in the interactome, so we ran the experiment 100 times with randomly added noise to the interactome and chose the top 400 nodes that appeared most frequently and removed any disconnected nodes. Additionally, we assessed the specificity of the network by assigning the input prize values to random nodes in the interactome and measuring the

frequency that each node appears. We repeated these experiments for a parameter grid and selected a network that 1) performed feature selection (i.e., did not include the entire input prize list), 2) was specific (as determined by the calculations using randomly assigned prizes), and 3) had a degree distribution that matched that of the input prize file. As C9ORF72 was not detected in the proteomics measurements, we forced C9ORF72 inclusion in the network by artificially assigning it a large prize. Network nodes were then sorted by subcellular location based on the Compartments database and plotted in Cytoscape (Binder et al., 2014).

Drosophila Screen

Drosophila orthologs of human DEGs were identified using DIOPT, and transgenic fly lines knocking-down or overexpressing these genes downstream of UAS sites for GAL4-specific modulation were obtained from the Bloomington Drosophila Stock Center (Hu et al., 2011). These modifiers were crossed to flies overexpressing the hexanucleotide repeat expansion (HRE) in the eye [GMR Gal4; UAS-(G4C2)30/CyO]. Progeny co-expressing both the HRE and putative modifier were collected within 24 hours of eclosion and aged at 25°C and compared to control flies of the same genetic background. A relative modification index, ranging from -4 to +4, was used to assess eye degeneration where -4 represented complete rescue and +4 represented no eye (Zhang et al., 2015). A score of 0 represents no effect of the tested modifier. Ommatidial structure, interommatidial bristles, necrosis, loss of pigmentation, and overall morphology of the eye were assessed during scoring. Only female flies were scored due to male flies displaying a higher degree of variability. All experimental modifiers were tested with 3 biological replicates with their eye degeneration scores averaged. If a fly cross failed to eclose, the subsequent score was marked 'lethal'. Selected strong enhancers and suppressors were retested with GMR Gal4; UAS-(G4C2)30/CyO as well as GMR Gal4 alone, at both 25°C and 29°C. At 15 days, representative female eyes were imaged using a Nikon SMZ1500 stereomicroscope and Lumenera INFINITY3-6UR 3.0 Megapixel camera and analyzed with Image-Pro Insight v9.

In some cases, a human candidate gene had multiple fly orthologs. For each human gene, a "weighted eye score" was calculated by taking the average of all corresponding fly orthologs, weighted by the ortholog scores as determined by the DRSC Integrative Ortholog Prediction Tool (https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl). Note that only moderate and high ranking orthologs were considered.

Drosophila Network Analysis

We categorized the genes that were tested in the Drosophila model into three groups: causal, compensatory, and non-contributory. For example, we reasoned that genes that were significantly upregulated in ALS and whose knockdown in fly suppressed or enhanced eye degeneration were likely causal or compensatory genes, respectively. Similarly, those that were significantly downregulated in ALS and were enhancers or suppressors of eye degenerations were likely causal or compensatory, respectively. Genes whose knockdown in the fly model showed little to no effect on eye degeneration were categorized as non-contributory.

Next, we used previously annotated directed interactions that were pulled from the ReactomeFiViz and KEGG databases (Kanehisa et al., 2017; Wu et al., 2014). The resulting directed network was composed of X nodes connected by Y directed edges. For any two proteins that were labeled as either causal or compensatory, we identified all directed paths of length at most 2. Next, we only considered paths that were concordant with our data by not allowing paths:

1) to contain genes that are not expressed in iMNs. This was defined by taking the top 70% of expressed gene transcripts across all 7 iMNs lines.

2) whose predicted effect on protein activity is discordant with measured protein expression. For instance, if A->B, but A is up in ALS and B is down in ALS, this edge is excluded from further analysis.

The resulting network was visualized by contracting proteins from the same complexes or protein families into single nodes (i.e. all ribosomal subunits are represented as one node), and the nodes were manually sorted by function and causal/compensatory role.

Statistical Analysis

Immunostaining: The boxplots shown in Figure B-1C are average results from quantified images of the respective immunostains in Figure B-1B. The healthy control donors (CTR) comprised of n = 3 independent iPSC lines, while the C9-ALS comprised of n = 4 *C9ORF72* repeat expansion donor iPSC lines. Total cells were quantified by nuclear staining with Hoechst 33258 in n = 9 sites across a well and percent positive cells for respective marker were calculated for each site. Average positive marker expression was then calculated for each well. Each marker immunostain was performed across independent well 3 times and respective average percent positive cells were obtained for each iPSC lines. All statistical analyses for percent SMI32, TuJ1, Map2a/b, GFAP and Nestin levels were performed using unpaired t test and the differences between CTR and C9-ALS groups were insignificant. Error bars represent SEM.

RM: Kaplan-Meier curves are constructed from individual survival times of cells from each well and survival of cohorts of cells from each well is compared to each other with survival or time-to-event analysis. Scripts written in R's survival package were used to generate cumulative risk of death curves and to perform cox proportional hazard analysis to assess the relative risk of death between the ALS and control motor neurons. For the supplementary table, Kaplan-Meier analysis revealed that the risk of death was not significantly high for C9-ALS lines as compared to CTR lines. In fact, control 00iCTR (n = 318 cells) line survived less well than 28iALS (hazard ratio (HR) is 0.6, p-value < 0.001, n = 264 cells), 30iALS (HR = 0.3, p-value < 0.001, n = 413 cells) and 52iALS (HR = is 0.54, p-value = 0.008, n = 75 cells). Similarly, control 25iCTR (n = 69 cells) line survived less well than 28iALS (HR = 0.47, p-value < 0.001, n = 264 cells), 30iALS (HR = 0.24, p-value < 0.001, n = 413 cells) and 52iALS (HR = 0.42, p-value = 0.001, n = 75 cells). Total n = 1188 cells; four experiments for 00iCTR and 25iCTR, three experiments for 28iALS, 30iALS and 52iALS.

RNA-Seq: Generalized linear models were used with negative binomial distribution to estimate fold change between ALS and controls samples for each gene. Wald test was performed for hypothesis testing, which is a one-sided test. Sample size n was 3 and 4 respectively for control and ALS.

Proteomics: Throughout Trans Proteome Pipeline (TPP) and OpenSWATH, a 1% FDR cutoff was employed in identification of transitions/peptides and in OpenSWATH matching to the peptide library. MAP DIA55 was used on MS2 normalized transition level data obtained from OpenSWATH. Transitions falling outside of 2 standard deviations were filtered out. Additional correlation filter of 0.2 was used to filter out any residual outliers. Intensities of the remaining transitions were summed for peptide, and then protein level quantification. Differential expression analysis of designated groups was performed by MAP DIA using analysis based on a Bayesian latent variable model with Markov random field prior. Output for differential expression included log2FC, confidence score, FDR and log(OddsofDifferentialExpression). Log2 fold changes were deemed significant if they had FDR at 1% or lower, a confidence score of .95 or above, a positive log(oddsDE) and an abs(log2FC) of .6 or above. For IPA analysis, the 924 differentially expressed proteins and their corresponding log2FC values were used, with analysis settings for reference set: Ingenuity Knowledge Bases, direct relationships, using all data sources, experimentally observed interactions and filtered for human genes in primary tissues and human cell lines. For pathway enrichment analysis, GOrilla was used (Eden et al., 2009). The DIA filtered list of 3,742 proteins was used as the background list for analysis of target sets. A p-value threshold of 10^-3 was used to determine enriched GO Biological Process terms.

ATAC-Seq: Differentially open sites were called using the DESeq2 pipeline with FDR <= 0.1.

Data integration: All GO enrichments were performed using a one-sided hypergeometric test implemented by GOrilla. Figure B-4A - Motif enrichments were calculated via HOMER which searches for de novo motif matches that are enriched in a set of foreground sequences relative to a given set of background sequences using a one-sided hypergeometric test. Figure B-5 - Enrichment of ALS-associated genes was calculated using a one-sided hypergeometric implemented using the hypergeometric module in Scipy v0.14. Enrichments of genes between omic assays were also calculated using a one-sided hypergeometric test implemented using the hypergeometric module in Scipy v0.14. For each pair of assays, the background was the set of genes that was detected in both assays.

Drosophila eye screen: flies were aged to 15 days after eclosion. 3 biological replicates were carried out per cross. 15 females flies were scored per cross. The average score of these 15 flies was taken as the average for one biological replicate. The average of all 3 biological replicates rounded to the nearest 0.5 of a point was used for the final rounded rough eye score.

B.6 References

Abel, O., Shatunov, A., Jones, A.R., Andersen, P.M., Powell, J.F., and Al-Chalabi, A. (2013). Development of a smartphone app for a genetics website: The amyotrophic lateral sclerosis online genetics database (ALSoD). J. Med. Internet Res. *1*, e18.

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. Nucleic Acids Res. *43*, D789-98.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., et al. (2009). NCBI GEO: Archive for high-throughput functional genomic data. Nucleic Acids Res. *41*, D991-5.

Bhinge, A., Namboori, S.C., Zhang, X., VanDongen, A.M.J., and Stanton, L.W. (2017). Genetic Correction of SOD1 Mutant iPSCs Reveals ERK and JNK Activated AP1 as a Driver of Neurodegeneration in Amyotrophic Lateral Sclerosis. Stem Cell Reports *8*, 856–869.

Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S.I., Schneider, R., and Jensen, L.J. (2014). COMPARTMENTS: Unification and visualization of protein subcellular localization evidence. Database bau012.

Bossis, G., Malnou, C.E., Farras, R., Andermarcher, E., Hipskind, R., Rodriguez, M., Schmidt, D., Muller, S., Jariel-Encontre, I., and Piechaczyk, M. (2005). Down-Regulation of c-Fos/c-Jun AP-1 Dimer Activity by Sumoylation. Mol. Cell. Biol. *25*, 6964–6979.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.

Brown, R., and Al-Chalabi, A. (2017). Amyotrophic Lateral Sclerosis. N. Engl. J. Med. *377*, 162–172.

Chodniewicz, D., and Klemke, R.L. (2004). Regulation of integrin-mediated cellular responses through assembly of a CAS/Crk scaffold. Biochim. Biophys. Acta - Mol. Cell Res. *1692*, 63–76.

Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res. *19*, 1553–1561.
Corbier, C., and Sellier, C. (2017). C9ORF72 is a GDP/GTP exchange factor for Rab8 and Rab39 and regulates autophagy. Small GTPases *8*, 181–186.
Coyne, A.N., Siddegowda, B.B., Estes, P.S., Johannesmeyer, J., Kovalik, T., Daniel, S.G., Pearson, A., Bowser, R., and Zarnescu, D.C. (2014). Futsch/MAP1B mRNA Is a Translational Target of TDP-43 and Is Neuroprotective in a Drosophila Model of Amyotrophic Lateral Sclerosis. J. Neurosci. *34*, 15962–15974.
Delic, V., Kurien, C., Cruz, J., Zivkovic, S., Barretta, J., Thomson, A., Hennessey, D., Joseph, J., Ehrhart, J., Willing, A.E., et al. (2018). Discrete mitochondrial aberrations in the spinal cord of sporadic ALS patients. J. Neurosci. Res. *96*, 1353–1366.
Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C.J., Gronostajski, R.M., and Anderson, D.J. (2006). The Transcription Factor NFIA Controls the Onset of Gliogenesis in the Developing Spinal Cord. Neuron *52*, 953–968.
DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.
Devlin, A.C., Burr, K., Borooah, S., Foster, J.D., Cleary, E.M., Geti, I., Vallier, L., Shaw, C.E., Chandran, S., and Miles, G.B. (2015). Human iPSC-derived motoneurons harbouring TARDBP or C9ORF72 ALS mutations are dysfunctional despite maintaining viability. Nat. Commun. *6*, 5999.
Dimos, J.T., Rodolfa, K.T., Niakan, K.K., Weisenthal, L.M., Mitsumoto, H., Chung, W., Croft, G.F., Saphier, G., Leibel, R., Goland, R., et al. (2008). Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. Science (80-.). *321*, 1218–1221.
Donnelly, C.J., Zhang, P.W., Pham, J.T., Heusler, A.R., Mistry, N.A., Vidensky, S., Daley, E.L., Poth, E.M., Hoover, B., Fines, D.M., et al. (2013). RNA Toxicity from the ALS/FTD C9ORF72 Expansion Is Mitigated by Antisense Intervention. Neuron *80*, 415–428.
Ebert, A.D., Yu, J., Rose, F.F., Mattis, V.B., Lorson, C.L., Thomson, J.A., and Svendsen, C.N. (2009). Induced pluripotent stem cells from a spinal muscular atrophy patient. Nature *457*, 277–280.
Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.
Edens, B.M., Yan, J., Miller, N., Deng, H.-X., Siddique, T., and Ma, Y.C. (2017). A novel ALS-associated variant in UBQLN4 regulates motor axon morphogenesis. Elife *6*, e25453.
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.
Fadista, J., Oskolkov, N., Hansson, O., and Groop, L. (2017). LoFtool: A gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics *33*, 471–474.

Farrer, L.A., Cupples, A.L., Kukull, W. a, Mayeux, R., Myers, R.H., Pericak-vance, M. a, Farrer, L. a, Cupples, L.A., Haines, J.L., Hyman, B., et al. (1997). Effects of Age , Sex , and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease. JAMA J. Am. Med. Assoc. *278*, 1349–1356.

Finkbeiner, S., Frumkin, M., and Kassner, P.D. (2015). Cell-based screening: Extracting meaning from complex data. Neuron *86*, 160–174.

Freibaum, B.D., Lu, Y., Lopez-Gonzalez, R., Kim, N.C., Almeida, S., Lee, K.H., Badders, N., Valentine, M., Miller, B.L., Wong, P.C., et al. (2015). GGGGCC repeat expansion in C9orf72 compromises nucleocytoplasmic transport. Nature *525*, 129–133.

Fujimori, K., Ishikawa, M., Otomo, A., Atsuta, N., Nakamura, R., Akiyama, T., Hadano, S., Aoki, M., Saya, H., Sobue, G., et al. (2018). Modeling sporadic ALS in iPSC-derived motor neurons identifies a potential therapeutic agent. Nat. Med. *24*, 1579–1589.

Fuller, H.R., Mandefro, B., Shirran, S.L., Gross, A.R., Kaus, A.S., Botting, C.H., Morris, G.E., and Sareen, D. (2016). Spinal Muscular Atrophy Patient iPSC-Derived Motor Neurons Have Reduced Expression of Proteins Important in Neuronal Development. Front. Cell. Neurosci. *9*, 506.

Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., Daniel, J.M.O., Ormond, K.E., et al. (2013). American College of Medical Genetics and Genomics ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. Genet. Med. *15*, 565–574.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res. *32*, D109-11.

Griffiths-Jones, S. (2005). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. *34*, D140-4.

Griffiths-Jones, S., Saini, H.K., Van Dongen, S., and Enright, A.J. (2008). miRBase: Tools for microRNA genomics. Nucleic Acids Res. *36*, D154-8.

Hanagasi, H.A., Giri, A., Kartal, E., Guven, G., Bilgiç, B., Hauser, A.K., Emre, M., Heutink, P., Basak, N., Gasser, T., et al. (2016). A novel homozygous DJ1 mutation causes parkinsonism and ALS in a Turkish family. Park. Relat. Disord. *29*, 117–120.

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E., Logroscino, G., Robberecht, W., Shaw, P., Simmons, Z., and van den Berg, L. (2017). Amyotrophic lateral sclerosis. Nat. Rev. Dis. Prim. *3*, 17071.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. *22*, 1760–1774.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hofmann, W.A., Arduini, A., Nicol, S.M., Camacho, C.J., Lessard, J.L., Fuller-Pace, F. V., and De Lanerolle, P. (2009). SUMOylation of nuclear actin. J. Cell Biol. *186*, 193–200.

Holewinski, R.J., Parker, S.J., Matlock, A.D., Venkatraman, V., and Van Eyk, J.E.

(2016). Methods for SWATH™: Data Independent Acquisition on TripleTOF Mass Spectrometers. Methods Mol. Biol. *1410*, 265–279.
Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., and Mohr, S.E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics *12*, 357.
Itan, Y., Shang, L., Boisson, B., Patin, E., Bolze, A., Moncada-Vélez, M., Scott, E., Ciancanelli, M.J., Lafaille, F.G., Markle, J.G., et al. (2015). The human gene damage index as a gene-level approach to prioritizing exome variants. Proc. Natl. Acad. Sci. *112*, 13615–13620.
Ji, Z., Degerny, C., Vintonenko, N., Deheuninck, J., Foveau, B., Leroy, C., Coll, J., Tulasne, D., Baert, J.L., and Fafeur, V. (2007). Regulation of the Ets-1 transcription factor by sumoylation and ubiquitinylation. Oncogene *26*, 395–406.
Jovičič, A., Mertens, J., Boeynaems, S., Bogaert, E., Chai, N., Yamada, S.B., Paul, J.W., Sun, S., Herdy, J.R., Bieri, G., et al. (2015). Modifiers of C9orf72 dipeptide repeat toxicity connect nucleocytoplasmic transport defects to FTD/ALS. Nat. Neurosci. *18*, 1226–1229.
Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361.
Kirk, J.A., Chakir, K., Lee, K.H., Karst, E., Holewinski, R.J., Pironti, G., Tunin, R.S., Pozios, I., Abraham, T.P., De Tombe, P., et al. (2015). Pacemaker-induced transient asynchrony suppresses heart failure progression. Sci. Transl. Med. *7*, 319ra207.
Koscielny, G., An, P., Carvalho-Silva, D., Cham, J.A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E., et al. (2017). Open Targets: A platform for therapeutic target identification and Validation. Nucleic Acids Res. *45*, D985-94.
Lagoutte, E., Villeneuve, C., Lafanechère, L., Wells, C.M., Jones, G.E., Chavrier, P., and Rossé, C. (2016). LIMK Regulates Tumor-Cell Invasion and Matrix Degradation Through Tyrosine Phosphorylation of MT1-MMP. Sci. Rep. *6*, 24925.
Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862-8.
Lee, S.W., Lee, M.H., Park, J.H., Kang, S.H., Yoo, H.M., Ka, S.H., Oh, Y.M., Jeon, Y.J., and Chung, C.H. (2012). SUMOylation of hnRNP-K is required for p53-mediated cell-cycle arrest in response to DNA damage. EMBO J. *31*, 4441–4452.
Lek, M., Karczewski, K.J., Minikel, E. V, Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
Lev, N., Barhum, Y., Lotan, I., Steiner, I., and Offen, D. (2015). DJ-1 Knockout augments disease severity and shortens survival in a mouse model of ALS. PLoS One *10*, e0117190.
Li, M.X., Gui, H.S., Kwan, J.S.H., Bao, S.Y., and Sham, P.C. (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. *40*, e53.
Li, M.X., Kwan, J.S.H., Bao, S.Y., Yang, W., Ho, S.L., Song, Y.Q., and Sham, P.C. (2013). Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genet. *9*, e1003143.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
Luan, Z., Liu, Y., Stuhlmiller, T.J., Marquez, J., and García-Castro, M.I. (2013). SUMOylation of Pax7 is essential for neural crest and muscle development. Cell. Mol. Life Sci. *70*, 1793–1806.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
McNeish, J., Gardner, J.P., Wainger, B.J., Woolf, C.J., and Eggan, K. (2015). From Dish to Bedside: Lessons Learned while Translating Findings from a Stem Cell Model of Disease to a Clinical Trial. Cell Stem Cell *17*, 8–10.
Milani, P., Escalante-Chong, R., Shelley, B.C., Patel-Murray, N.L., Xin, X., Adam, M., Mandefro, B., Sareen, D., Svendsen, C.N., and Fraenkel, E. (2016). Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. Sci. Rep. *6*, 25474.
Moller, A., Bauer, C.S., Cohen, R.N., Webster, C.P., and De Vos, K.J. (2017). Amyotrophic lateral sclerosis-associated mutant SOD1 inhibits anterograde axonal transport of mitochondria by reducing Miro1 levels. Hum. Mol. Genet. *26*, 4668–4679.
Ng, B., White, C.C., Klein, H.U., Sieberts, S.K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D.A., et al. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat. Neurosci. *20*, 1418–1426.
Ng, S.Y., Soh, B.S., Rodriguez-Muela, N., Hendrickson, D.G., Price, F., Rinn, J.L., and Rubin, L.L. (2015). Genome-wide RNA-Seq of Human Motor Neurons Implicates Selective ER Stress Activation in Spinal Muscular Atrophy. Cell Stem Cell *17*, 569–584.
Nizzardo, M., Simone, C., Dametti, S., Salani, S., Ulzi, G., Pagliarani, S., Rizzo, F., Frattini, E., Pagani, F., Bresolin, N., et al. (2015). Spinal muscular atrophy phenotype is ameliorated in human motor neurons by SMN increase via different novel RNA therapeutic approaches. Sci. Rep. *5*, 11746.
Paez-Colasante, X., Figueroa-Romero, C., Sakowski, S., Goutman, S., and Feldman, E. (2015). Amyotrophic lateral sclerosis: Mechanisms and therapeutics in the epigenomic era. Nat. Rev. Neurol. *11*, 266–279.
Palmesino, E., Rousso, D.L., Kao, T.J., Klar, A., Laufer, E., Uemura, O., Okamoto, H., Novitch, B.G., and Kania, A. (2010). Foxp1 and Lhx1 coordinate motor neuron migration with axon trajectory choice by gating reelin signalling. PLoS Biol. *8*, e1000446.
Paré, B., Lehmann, M., Beaudin, M., Nordström, U., Saikali, S., Julien, J.P., Gilthorpe, J.D., Marklund, S.L., Cashman, N.R., Andersen, P.M., et al. (2018). Misfolded SOD1 pathology in sporadic Amyotrophic Lateral Sclerosis. Sci. Rep. *8*, 14223.
Parker, S.J., Venkatraman, V., and Van Eyk, J.E. (2016). Effect of peptide assay library size and composition in targeted data-independent acquisition-MS analyses. Proteomics *16*, 2221–2237.
Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic

Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. *9*, e1003709.

Piñero, J., Bravo, Á., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45*, D833-9.

Prudencio, M., Belzil, V. V., Batra, R., Ross, C.A., Gendron, T.F., Pregent, L.J., Murray, M.E., Overstreet, K.K., Piazza-Johnston, A.E., Desaro, P., et al. (2015). Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. Nat. Neurosci. *18*, 1175–1182.

Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.

Robberecht, W., and Philips, T. (2013). The changing scene of amyotrophic lateral sclerosis. Nat. Rev. Neurosci. *14*, 248–264.

Sances, S., Bruijn, L.I., Chandran, S., Eggan, K., Ho, R., Klim, J.R., Livesey, M.R., Lowry, E., Macklis, J.D., Rushton, D., et al. (2016). Modeling ALS with motor neurons derived from human induced pluripotent stem cells. Nat. Neurosci. *19*, 542–553.

Sanfilippo, C., Longo, A., Lazzara, F., Cambria, D., Distefano, G., Palumbo, M., Cantarella, A., Malaguarnera, L., and Di Rosa, M. (2017). CHI3L1 and CHI3L2 overexpression in motor cortex and spinal cord of sALS patients. Mol. Cell. Neurosci. *85*, 162–169.

Sareen, D., Ebert, A.D., Heins, B.M., McGivern, J. V., Ornelas, L., and Svendsen, C.N. (2012). Inhibition of apoptosis blocks human motor neuron cell death in a stem cell model of spinal muscular atrophy. PLoS One *7*, e39113.

Sareen, D., O'Rourke, J.G., Meera, P., Muhammad, A.K.M.G., Grant, S., Simpkinson, M., Bell, S., Carmona, S., Ornelas, L., Sahabian, A., et al. (2013). Targeting RNA foci in iPSC-derived motor neurons from ALS patients with a C9ORF72 repeat expansion. Sci. Transl. Med. *5*, 208ra149.

Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.

Selvaraj, B.T., Livesey, M.R., Zhao, C., Gregory, J.M., James, O.T., Cleary, E.M., Chouhan, A.K., Gane, A.B., Perkins, E.M., Dando, O., et al. (2018). C9ORF72 repeat expansion causes vulnerability of motor neurons to Ca2+-permeable AMPA receptor-mediated excitotoxicity. Nat. Commun. *9*, 347.

Shelley, B.C., Gowing, G., and Svendsen, C.N. (2014). A cGMP-applicable Expansion Method for Aggregates of Human Neural Stem and Progenitor Cells Derived From Pluripotent Stem Cells or Fetal Brain Tissue. J. Vis. Exp.

Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. *111*, E5593-601.

Shi, Y., Lin, S., Staats, K.A., Li, Y., Chang, W.H., Hung, S.T., Hendricks, E., Linares, G.R., Wang, Y., Son, E.Y., et al. (2018). Haploinsufficiency leads to neurodegeneration in C9ORF72 ALS/FTD human induced motor neurons. Nat. Med. *24*, 313–325.

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452-7.

Sivadasan, R., Hornburg, D., Drepper, C., Frank, N., Jablonka, S., Hansel, A., Lojewski, X., Sterneckert, J., Hermann, A., Shaw, P.J., et al. (2016). C9ORF72 interaction with cofilin modulates actin dynamics in motor neurons. Nat. Neurosci. *19*, 1610–1618.

Skibinski, G., Nakamura, K., Cookson, M.R., and Finkbeiner, S. (2014). Mutant LRRK2 Toxicity in Neurons Depends on LRRK2 Levels and Synuclein But Not Kinase Activity or Inclusion Bodies. J. Neurosci. *34*, 418–433.

Solomon, B.D., Nguyen, A.-D., Bear, K.A., and Wolfsberg, T.G. (2013). Clinical Genomic Database. Proc. Natl. Acad. Sci. *110*, 9851–9855.

Soltis, A.R., Motola, S., Vernia, S., Ng, C.W., Kennedy, N.J., Dalin, S., Matthews, B.J., Davis, R.J., and Fraenkel, E. (2017). Hyper- and hypo- nutrition studies of the hepatic transcriptome and epigenome suggest that PPARα regulates anaerobic glycolysis. Sci. Rep. *7*, 174.

Song, F., Chiang, P., Wang, J., Ravits, J., and Loeb, J.A. (2012). Aberrant neuregulin 1 signaling in amyotrophic lateral sclerosis. J. Neuropathol. Exp. Neurol. *71*, 104–115.

Suttkus, A., Morawski, M., and Arendt, T. (2016). Protective Properties of Neural Extracellular Matrix. Mol. Neurobiol. *53*, 73–82.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science (80-.). *337*, 64–69.

Teo, G., Kim, S., Tsou, C.C., Collins, B., Gingras, A.C., Nesvizhskii, A.I., and Choi, H. (2015). MapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. J. Proteomics *129*, 108–120.

Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput. Biol. *12*, e1004879.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science (80-.). *347*, 1260419.

Vazquez-Arango, P., Vowles, J., Browne, C., Hartfield, E., Fernandes, H.J.R., Mandefro, B., Sareen, D., James, W., Wade-Martins, R., Cowley, S.A., et al. (2016). Variant U1 snRNAs are implicated in human pluripotent stem cell maintenance and neuromuscular disease. Nucleic Acids Res. *44*, 10960–10973.

Wainger, B.J., Kiskinis, E., Mellin, C., Wiskow, O., Han, S.S.W., Sandoe, J., Perez, N.P., Williams, L.A., Lee, S., Boulting, G., et al. (2014). Intrinsic membrane hyperexcitability of amyotrophic lateral sclerosis patient-derived motor neurons. Cell Rep. *7*, 1–11.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wei, F., Schöler, H.R., and Atchison, M.L. (2007). Sumoylation of Oct4 enhances its stability, DNA binding, and transactivation. J. Biol. Chem. *282*, 21551–21560.

Wilson, J.M. (2005). Conditional Rhythmicity of Ventral Spinal Interneurons Defined by Expression of the Hb9 Homeodomain Protein. J. Neurosci. *25*, 5710–5719.
Wroe, R., Wai-Ling Butler, A., Andersen, P.M., Powell, J.F., and Al-Chalabi, A. (2008). ALSOD: The amyotrophic lateral sclerosis online database. Amyotroph. Lateral Scler. *9*, 249–250.
Wu, G., Dawson, E., Duong, A., Haw, R., and Stein, L. (2014). ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. F1000Research *3*, 146.
Xu, Z., Poidevin, M., Li, X., Li, Y., Shu, L., Nelson, D.L., Li, H., Hales, C.M., Gearing, M., Wingo, T.S., et al. (2013). Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. Proc. Natl. Acad. Sci. *110*, 7778–7783.
Zhang, K., Donnelly, C.J., Haeusler, A.R., Grima, J.C., Machamer, J.B., Steinwald, P., Daley, E.L., Miller, S.J., Cunningham, K.M., Vidensky, S., et al. (2015). The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. Nature *525*, 56–61.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.