From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

INTEGRATIVE OMICS DATA ANALYSIS TO DISCOVER NOVEL SIGNATURES IN COMPLEX DISEASES

Lu Pan

潘璐

Karolinska Institutet

Stockholm 2023

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2023

ISBN 978-91-8016-833-5
Cover illustration: A futuristic illustration of multi-omics. Descriptive text suggested by ChatGPT and AI imaging by Midjourney.

Integrative omics data analysis to discover novel signatures in complex diseases

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Lu Pan

The thesis will be defended in public at the Atrium, Nobels väg 12B, Solna, 9:00AM on the 31st of March, 2023.

Principal Supervisor:
Trung Nghia Vu
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor(s):
Fang Fang
Karolinska Institutet
Institute of Environmental Medicine

Xia Shen
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Yudi Pawitan
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Francesca Cordero
University of Turin
Department of Computer Science

Examination Board:
Mika Gustafsson
Linköping University
Department of Physics, Chemistry and Biology

Åsa Johansson
Uppsala University
Department of Immunology, Genetics and Pathology

Keith Humphreys
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

To my mother and my grandparents

ABSTRACT

Apart from diseases caused by the defect of a single gene, most diseases are highly complex and are usually caused by a combination of biological and environmental factors. In the biological context, cellular processes are often tightly connected across molecular layers of the central dogma of biology, and the examination of a single layer would not be sufficient to address disease pathology, therefore, conclusions drawn can be limited. Combining biological observations from multiple layers or angles would greatly broaden our perspectives on the disease in concern and may lead to novel discoveries which would not be possible to deduce from a single-omics perspective. In this thesis, we focused on the method development for single-cell transcriptomics to address the prime bias problem introduced by the new droplet-based technologies; integrative omics discovery of genomic signatures specific to different brain regions in normal individuals; as well as the utilization of multiple omics to identify potential biomarkers specific to amyotrophic lateral sclerosis (ALS) disease prognosis and diagnosis.

Research has been revolutionized with the advent of single-cell omics technologies in the past few decades and new methods and tools have also been developed to accommodate such scientific accelerations. These innovations however posed new challenges and could potentially introduce bias and unforeseeable circumstances if left unaddressed. Specifically, to resolve the prime-based problem introduced by the current popular droplet-based single-cell sequencing technologies which may lead to bias quantification, in Study I, we presented a novel transcript quantification tool for droplet-based single-cell RNA-Sequencing (scRNA-Seq) technologies and benchmarked our tool with other popular transcript and gene quantification tools. Our tool outperformed currently popular tools in terms of transcript- and gene-level quantifications.

In Study II, we investigated the association of splicing variants with the genetic patterns from different regions of the brain in normal individuals to identify quantitative trait loci (QTL) associated with ratios of isoform expression in genes. We carried out genome-wide association studies (GWAS) on isoform ratios from 13 brain regions and identified isoform-ratio QTL (irQTL) specific to each brain region, and their associated traits which could have been missed by expression QTL derived from gene expressions.

We further looked into the utilization of proteomics and genomics data for ALS disease in Study III to understand disease pathology from multiple perspectives, and to identify potential protein biomarkers and protein QTL (pQTL) specific to different stages of the disease and tissue sites. In terms of proteomics, for each tissue site, we identified potential protein biomarkers specific to disease prognosis, survival of ALS patients, the functional decline among ALS patients, and longitudinal changes after disease diagnosis. In terms of integrative omics, we performed GWAS of protein expressions with genotyping data and identified tissue-site-specific pQTL signatures for ALS patients.

All in all, our studies showed efforts in developing a single-cell transcript quantification tool to address potential bias problems with improved performance; identifying novel irQTL signatures specific to various brain regions using an integrative omics approach; and also discovering potential protein and genetic signatures for different tissues sites and pathological stages in ALS disease using multiple omics. We hope our work could potentially enhance the research process in various omics in terms of methods development and the novel signatures could act as valuable resources for fostering further research ideas and potential experimental validations.

LIST OF SCIENTIFIC PAPERS

I. **Lu P**, Huy QD, Yudi P, Trung Nghia V.
Isoform-level quantification for single-cell RNA sequencing.
Bioinformatics. 2022; 38(5).

II. **Lu P**, Chenqing Z, Zhijian Y, Yudi P, Trung Nghia V, Xia S.
Hidden genetic regulation of human complex traits via brain isoforms.
Phenomics (In press).

III. **Lu P**, Christina S, Caroline I, Åsa H, Jose L, Trung Nghia V, Yudi P, Abbe U, Solmaz Y, John A, Emily J, Aniko L, Yan C, Kristin S, Rayomand P, Fredrik P, Caroline G, Anders M, Fang F.
Protein biomarkers in risk and prognosis of amyotrophic lateral sclerosis.
Manuscript in preparation.

SCIENTIFIC PAPERS NOT INCLUDED IN THE THESIS

1. **Lu P**, Shaobo S, Roman T, Weiyuan L, Zehuan L, Hangyu S, Qishuang C, Xiaolu Z, Xuexin L.
   HTCA: a database with an in-depth characterization of the single-cell human transcriptome.
   Nucleic Acids Research. 2022; 51(D1).

2. Solmaz Y, Christina S, Can C, Anikó L, **Lu P**, Fredrik P, *et al.*
   T cell responses at diagnosis of amyotrophic lateral sclerosis predict disease progression.
   Nature Communications 2022; 13(6733).

3. Suk PC, Salvatore A, Kah-hoe PC, **Lu P**.
   System and method for classifying cancer patients into appropriate cancer treatment groups and compounds for treating the patient.
   US Patent Application Publication. 2019; Patent No. 16/768,001.

4. Joo GY, Martin W, Pavanish K, **Lu P**, Su LP, Fauziah A, *et al.*
   The Extended Polydimensional Immunome Characterization (EPIC) web-based reference and discovery tool for cytometry data.
   Nature Biotechnology. 2020; 38(6).

5. Chun JL, Yun HL, **Lu P**, Liyun L, Camillus C, Martin W, *et al.*
   Multidimensional analyses reveal distinct immune microenvironment in hepatitis B virus-related hepatocellular carcinoma.
   Gut. 2019; 68(5).

6. Valerie C, Yun HL, **Lu P**, Nurul JMN, Chun JL, Camillus C, *et al.*
   Immune activation underlies a sustained clinical response to Yttrium-90 radioembolisation in hepatocellular carcinoma.
   Gut. 2019; 68(2).

CONTENTS

LIST OF ABBREVIATIONS

Amyotrophic lateral sclerosis	ALS
Alternating expectation-maximization	AEM
Benjamin-Hochberg	BH
Blood-brain barrier	BBB
Body mass index	BMI
Central nervous system	CNS
Cerebrospinal fluid	CSF
Equivalence classes	eqClass
Expectation step	E step
Expression quantitative trait loci	eQTL
False-discovery rate	FDR
Generalized estimating equation	GEE
Genome-wide association studies	GWAS
Genotype-tissue expression	GTEx
Human immunodeficiency virus	HIV
Insertions and deletions	Indels
Isoform-ratio quantitative trait loci	irQTL
Long non-coding RNAs	lncRNAs
Mass cytometry	CyTOF
Mass spectrometry	MS
Maximization step	M Step
Mendelian randomization	MR
Methylation quantitative trait loci	mQTL
Motor neuron disease	MND
Motor neurons	MNs
Next generation sequencing	NGS
Normalized Protein expression	NPX
Olink proximity extension assay	PEA
Poly-protein risk score	PRS
Polymerase chain reaction	PCR
Principle components	PCs
Protein quantitative trait loci	pQTL
Quantitative trait loci	QTL
Revised amyotrophic lateral sclerosis functional rating scale	ALSFRS-R
RNA-sequencing	RNA-Seq
Same transcript cluster	TC
Single-cell RNA-Sequencing	scRNA-Seq
Single-nucleotide polymorphisms	SNPs
Splicing quantitative trait loci	sQTL
Stratified linkage disequilibrium score regression	S-LDSC
Transcript compatibility count	TCC
Transcript per million	TPM
Uk Biobank	UKB
Unique molecular identifier	UMI
Uppsala multidisciplinary center for advanced computational science	UPPMAX
Whole-genome sequencing	WGS

1 INTRODUCTION

1.1 OMICS STUDIES

The burgeoning of large-scale high-throughput biotechnologies in the past few decades has enabled us to observe complex molecular events with heightened dimensionalities and precisions. We are now able to capture snapshots of biological phenomena across multiple layers following the central dogma of biology[3] with the aid of multiple omics technologies, such as genomics, epigenetics, transcriptomics, and proteomics. The suffix "-omics", concatenated to the molecular terms in the central dogma, is defined as "a comprehensive or global assessment of a set of molecules"[4,5], with each omics type unveiling a single layer of molecular information from a complex biological system. Biological measurements were further revolutionized and increased in dimensionalities in recent decades through the transformation of these technologies into single-cell technologies. As biological events are highly complex processes involving cellular interactions between molecules, cells, or tissues[6], measuring a single molecular modality would not be able to comprehensively assess the entire spectrum of complex interactions in most human diseases[7,8], especially for complex diseases when the cause of a disease is usually not attributed to the abnormality of a single effector gene[8-10]. Many underlying biological mechanisms of diseases are still largely unknown due to the complexity of the biological systems and the transformation of mainstream research into multi-omics perceptions is imminent. Adapting a multi-omics approach, by examining several omics at a time, would provide multiple perspectives for the same problem so that we could dissect, connect, and trace highly intertwined events across molecular layers to reveal underlying biological mechanisms missed by a single omics approach.

1.2 POTENTIAL LIMITATIONS IN OMICS TECHNOLOGIES

The accelerated progress in omics development has urged perpetual transformations and developments of methods for analyzing these new data forms with increased dimensionalities. Novel methods and tools have been developed quickly to acclimatize to such rapid changes. However, with new technologies come new challenges. Due to the nature of the new techniques which produced heterogeneous outputs with varying natures[11], many challenges remained unresolved in both the technical and the analytical aspects of the omics studies[6,11-14]. In the area of single-cell transcriptomics, droplet-based technologies such as 10X Genomics, tend to produce libraries with strong biases of read distributions at the transcript prime ends (i.e., 3' or 5' ends)[15], generating highly-similar read statistics between similar isoforms, thus creating problems for the precise quantification of transcripts. Efforts have been made to quantify isoforms at the single-cell level[16-19], yet precise quantification of transcripts for prime-bias single-cell droplet-based techniques remained a challenge.

1.3 OPPORTUNITIES FOR IMPROVEMENTS

This thesis aims to address potential problems introduced by the current omics technologies, especially in the attempt to quantify isoforms in single-cell droplet-based techniques and to validate existing and discover novel molecular signatures using multi-omics approaches in

normal brain tissues and complex diseases. In the later part of the thesis, we focus on ALS, which is a complex, fatal, and progressive neurodegenerative disease caused by the gradual deterioration of motor neurons with a mean survival period of three to five years after disease onset[20]. Due to its highly complex nature in disease pathways, ALS remains uncured and no effective treatment strategies have been discovered[21]. We aimed to discover novel biomarkers specific to the disease in terms of diagnosis, prognosis, survival, and patient functional declination. The overall goal of the thesis is to address the prime-biased problem in isoform quantification in single-cell transcriptomics and to utilize multiple omics to discover novel signatures and disease phenotypes that could potentially be used in the clinical research for diagnosis, prognosis, or drug-target design for ALS or other neuron-related diseases.

2 LITERATURE REVIEW

In this review, we gave an overview of omics types and their current developments, specifically, in the area of genomics, transcriptomics, and proteomics. We have also looked into the current development of single-cell isoform quantification. In the end, we summarized biomarker discovery, and disease subtyping at different omics levels in ALS.

2.1 OMICS TYPES AND THEIR DEVELOPMENTS

To vividly depict the study of different layers in biology, the characterization of biosystems or organisms at different biomolecular levels is known as an omics cascade[5,22]. The history of omics cascade could be traced back to the first successful attempt to isolate DNA by Friedrich Miescher[23] in 1869, and in 1953 when DNA double helix structure was first discovered by James Watson and Francis Crick[24]. Following the success of helix discovery, the ability to determine the sequence of genetic barcodes in the DNA came into realization with the development of the "PLUS AND MINUS" method by Sanger[25] in 1975, as well as the parallel developments of the Maxam-Gilbert method[26] and the chain-terminating Sanger method (or dideoxy method)[27] in 1977. These are known as the first generation of sequencing methods[28]. First-generation sequencing methods opened doors to investigate genomic sequences of organisms, however, drawbacks such as low efficacy in volume have motivated scientists to seek further improvements in sequencing technologies. Improvements came via the development of second-generation sequencing, which is also known as the Next Generation Sequencing (NGS)[23,29] in the 1990s. NGS allows simultaneous massively parallel sequencing of millions of DNA fragments at the same time, while maintaining cost-efficient characteristics and speed-wise being much faster than the first-generation technology. Despite NGS as a cutting-edge technology being widely used to study genome biology, there are issues such as polymerase chain reaction (PCR) errors during DNA amplification that might contribute to false discovery and interpretation in genetic analyses[30]. In that case, third-generation sequencing was initiated in 2008 to circumvent the hurdles by retaining higher integrity of DNA during sequencing without having to break them down into fragments before amplification[23,29,31]. The technology will significantly enhance the overall quality of genome assemblies[32,33]. These sequencing technologies serve as the basis of molecule sequencing and amplification for most omics.

2.1.1 Genomics

Genomics refers to a comprehensive assessment of the genome[34]. Investigations into genomics allow researchers to gain a deeper understanding of the genetic makeups of organisms, especially with the accomplishment of the assembly of the entire human genome, the Human Genome Project[35,36] in 2003, which is one of the greatest scientific achievements in genomics. As the cost of sequencing has drastically decreased over the past decades[37], population-based genotyping such as the UK Biobank[38] (UKB) was constructed to enable the study of genetic variations across thousands of people. GWAS, a methodology developed to identify genetic variants contributing to the strong associations between certain genotypes and corresponding

phenotypes[39], has produced an encyclopedia of genetic loci associated with specific phenotypes, including disease traits[40]. With the advent of other omics technologies, the association analysis can be done across the molecular layers of the central dogma[40] to deduce casual gene-to-transcript, gene-to-proteins, etc. relationships. These association studies across complex molecular events would not have been made possible with the use of only genomics.

2.1.2 Transcriptomics

Transcriptomics is the study of the transcriptional events present in a cell, tissue, or organism[41] using high-throughput sequencing technologies such as the RNA-Sequencing (RNA-Seq) technologies[42]. The utilization of RNA-Seq can provide an in-depth understanding of the fundamental mechanisms of the complex gene expression networks in highly-connected biological systems such as cancers[43]. In merely less than two decades, after the first single-cell sequenced by Tang et al.[44,45] back in 2009, transcriptomics development transformed itself from bulk sequencing of cellular populations to single-cell transcriptomics and spatially-resolved transcriptomics. We are now able to detect expression profiles of individual cells, reveal complex and novel cell populations that can never be imagined at the bulk level, uncover regulatory relationships between genes and cells, and carry out trajectory tracing of distinct cell lineages in cellular developments[46]. Cellular subtypes are now identifiable with high precisions via the cellular level transcript phenotyping with a wide range of single-cell transcriptomics technologies. As technologies are evolving, so are their downstream handling methods. Many conventional bioinformatics approaches used for handling bulk RNA-Seq data are no longer suitable to apply to single-cell data as the dimensionality of the data has evolved to higher dimensions, and the nature of these new data forms is different due to the variations in single-cell sequencing techniques. This still poses challenges in scRNA-Seq analyses across a variety of RNA-Seq protocols, such as the prime-biased problem in droplet-based technologies, causing difficulties in isoform quantifications. In recent years, multi-modal experimental techniques have been invented to study multiple modalities on a single cell[6]. This has advanced single-cell omics to the next level, beyond the need for integrative analysis of omics data from different cells, and reduced cell-to-cell variations during integration between different modalities. There are also challenges in this advanced technology, including its high cost which limits large-scale multi-modal studies, low coverage, and limited choices of multi-modal layers[6].

2.1.3 Proteomics

Proteomics aims to assess phenotypically, the diverse characteristics of proteins, including their structures, functions, interactions, etc[47,48]. Up-to-date, high-throughput profiling across a large number of protein types became possible with technologies such as Olink™ at the bulk proteomics level, and antibodies using mass cytometry (CyTOF) at the single-cell level, with CyTOF developed based on the traditional mass spectrometry (MS) technique. Proteomics technological developments have helped in biomarker identifications for many diseases with promising clinical applications such as disease diagnosis or prognosis in many proteomics studies[49-55]. Of all tissues, many studies focused on plasma as the main discovery medium for

biomarkers due to its comprehensiveness in human proteome and the prioritization for an eventual blood test of biomarkers[56]. This created new complexities in biomarker discoveries as huge varieties of proteins are present in the blood plasma and protein products varies from cell to cell. The limitations of the current technologies in quantifying and identifying proteins specific to different diseases are still a challenge, especially for complex diseases. Some resort to the use of other sampling sites for biomarker identification, including cerebrospinal fluid (CSF), brain extracts, etc[56,57]. However, samples from such sampling sites are sparse, and difficult to procure large sample numbers for many studies. As proteins are translational products of genes, there are essentially complex linkages between the genotype and the phenotype (proteins). What determines the architecture of a protein depends on multiple genetic and environmental factors, and the co-existence of other omics technologies fosters strong support for multi-omics integration of proteomics with other omics.

2.2 ISOFORM QUANTIFICATION IN SCRNA-SEQ

The significant progress in the development of sequencing technology and computational methods has brought the exploration of RNA expression measurement and quantification from bulk to single-cell level. A critical question related to RNA expression is to investigate alternative splicing patterns, which is a regulation mechanism in most genes. The quick advancements in transcriptomics technologies have introduced varying technical protocols with considerable technical and causing-leading downstream analysis limitations[42,45,58]. Popular droplet-based technologies such as 10X Genomics, which focus on prime-end sequencing protocols, tend to have a non-uniform distribution of reads across the cDNA body, especially with reduced reads mapped to positions further away from the primed regions. This prime-bias phenomenon created further challenges in the estimation of isoforms at the single-cell level. For full-length scRNA-Seq protocols which have no prime-bias limitations, several methods have been developed to estimate isoform expression[17-19]. Even though an attempt has been made to quantify splicing events in prime-based scRNA-Seq data[16], the ability to quantify isoform expression while modeling the prime-bias problem in single-cell droplet-based technologies is still a challenge. Addressing such limitations would improve the accuracy of isoform quantifications in droplet-based technologies.

2.3 GENETIC REGULATIONS IN THE MOST COMPLEX ORGAN OF THE HUMAN BODY – THE BRAIN

Gene regulation is the basis of the entire biological activities. It is not restricted to a single or isolated gene but is controlled by comprehensive regulatory networks. With the increasing development in high-throughput biochemical assay and understanding of the complex gene regulation mechanisms, omics technology has become an extremely powerful tool for the investigation of QTL in complex human tissues, such as the most variable region of the human body, the brain. With the emergence of different omics, the discovery of QTL has been leveraged by the multi-omics approach. For instance, comparing gene expressions (here it acts as the phenotype) with genetic variants (the genotype) would give rise to expression QTL (eQTL), which are genomic loci with effects on gene expression phenotype[59]; comparing

splicing patterns of genes (phenotype) with genetic variants would lead to the identification of splicing QTL (sQTL), which are genomic loci with effects on alternative splicing events[60]; comparing protein expression (phenotype) with genetic variants would determine pQTL, which are genomic loci with effects on protein expression[61]. Through multi-omics QTL analysis, we are now able to identify causal genotype-phenotype relationships between different molecular layers. The QTL analysis has also been extensively applied to study the most complex tissue in the human body, the brain tissue. The best example of its application in biological findings is the Genotype-Tissue Expression (GTEx) project[60], which is a pioneering project to massively phenotype gene expression and discover tissue-specific QTL across various human tissues, including 13 regions across the brain tissue. As the largest human tissue transcriptome expression database to date, many studies rely on this database for further genomics and transcriptomics studies. Apart from GTEX, many other studies have also made outstanding contributions in revealing genetic variations among a diverse set of human tissues. In 2019, the largest eQTL and sQTL map specific to human prenatal brain development was published and unearthed many potential risk loci specific to neurodevelopmental and neuropsychiatric disorders[62]. Zhang *et al.* identified sQTL with regional-specificity across different brain regions, suggesting the importance of regional variation of genetic variants in controlling splicing patterns[63]. In a recent study, Wang *et al.* identified methylation QTL (mQTL) with substantial effects on disease progression for Alzhermier's disease[64]. In addition, Jack H. *et al.* in 2021 identified potential risk loci in ALS patients based on the spinal cord transcriptome profiles of the diseased patients[65]. The progress in brain-specific QTL discoveries in multi-omics settings provided potential guidance for further variant identifications in many other diseases.

2.4 AMYOTROPHIC LATERAL SCLEROSIS

ALS is a non-curable neurogenerative disease involving the gradual degeneration of motor neurons in the central nervous system (CNS) with a median survival of 2-4 years after disease onset[20,21]. The disease is often highly complicated with genetic, environmental, lifestyle, and time factors, with heterogeneous clinical representations and disease phenotypes overlapping with that of other diseases, complicating our overall understanding of the disease, as well as subsequent strategies to the diagnosis, prognosis, and effective treatments for the disease.

2.4.1 Epidemiology

ALS is considered a rare disease with a standardized global incidence rate of only 1.68 per 100,000 person-years with varying regional statistics[21]. The incidence rate is higher in Europe and North America with a range of 1.71 to 1.89 per 100,000 person-years, and lower in Asia with a range of 0.83 to 0.94 per 100,000 person-years[21]. In terms of age, ALS incidence increases with age[21] and peaks at the age of 60 to 79 years[21]. The incidence happened to demonstrate an upward trend in recent decades which could have been a cause of social advancements such as improved diagnosis in the clinical settings[21]. Age, sex, and genetic factors contributing to disease incidence are often intertwined. For sex, the standardized incidence rate is 1.35 for the male-to-female ratio, with higher disease inheritance from mother to daughter, and lower disease onset age for men with *C9orf72* mutant gene compared to

women[21]. In terms of disease phenotypes, women above 60 years are more prone to bulbar disease onset than men, whereas men with age less than 60 years often possess classical disease phenotypes, which is the weakening of muscles starting from the limbs with degenerations in both upper and lower motor neurons (MNs)[21].

2.4.2 Disease Causes

Even though the causes of ALS are often highly complex, the causes could still be generally classified into two classes. Around 15% of ALS patients are caused by familial genetic inheritance (hence termed familial ALS) and the rest of the 85% are sporadic with unknown causes[21,66]. Genetic components causing ALS are at large, attributed to the inheritance of single-gene mutations[21,67]. Up-to-date, over 40 ALS-related genes have been discovered, with the most common gain-of-function mutation, *C9orf72* expansions, which paralyzes RNA metabolism[21,68]. Mutations in these genes increased the toxicity of the relevant biological environments by producing or lost-to-produce protein products destructing or maintaining the normal molecular pathways in the system through the gain-of-function or loss-of-function mutations[21]. Even though these ALS-related genes have been discovered, some of the mutation patterns of these genes were also found in other diseases, thus complicating the overall ALS diagnosis[21,69-71]. For sporadic ALS with no family history of ALS, genetic risk variants could also be found in these patients and can be oligogenic or polygenic[72] with earlier disease onset compared to those with a single or no such variants[21,72]. This may be due to the misclassification of these patients into sporadic cases due to limited family history information[21]. Other than genetics itself, environmental, lifestyle, and time factors, as well as the interactions between them, are also critical in contributing to the onset of the disease[20,21,73].

2.4.3 Clinical Representations and Disease Phenotypes

ALS is highly heterogeneous in terms of clinical representations and outcome phenotypes due to the combination of disease-causing factors and their corresponding combined outcomes. The disease at large often caused dysfunction in the MNs and could be in either-or or both upper and lower MNs[21,74]. This degeneration cascaded into chain reactions depending on the affected zones and will lead to muscles involved in voluntary movements with diverse clinical representations[21,74]. Cognitive and behavioral changes could also show up in 35-50% of ALS patients[21]. Symptoms include and not limited to speech difficulties, memory deficits, irritations in behaviors, depression, sleeping disorders, etc[21,74]. Some of these representations fulfilled the diagnostic criteria for other similar diseases such as frontotemporal dementia[21] and further complicated the process of understanding the disease.

2.4.4 Potential Genetic Therapies and Protein Biomarkers

New clinical trials have been designed to target ALS-associated genes with gain-of-function mutations[21]. Just like the famous genetically-modified Flavr Savr tomato[75], one of the strategies is to design antisense RNAs to target and complementary bind to the RNA molecules (pre- or matured-RNAs) to prevent further translating these mutant RNAs into toxic protein products hijacking its targeted molecular environments[21]. Potential prognostic and diagnostic

biomarkers are also one of the targeted approaches for clinical use. The neurofilaments[76] were robustly been reported in many studies as potential biomarkers for survival, disease diagnosis, phenotypes, etc[21,66,77,78]. As proteins are ubiquitously present in most body tissues with the need for normal tissue functioning and communications with other body components, more proteins could be investigated and targeted to assess their relevance in disease diagnosis, prognosis, etc[66].

2.4.5 Status-Quo

Up to today, there is no cure or effective treatment for ALS. However, the advancements in technologies and social infrastructures with better healthcare systems and more complete patient information have greatly facilitated our research into forging better understandings of the underlying disease mechanisms and greater improvements in the clinical settings with enhanced diagnostic criteria. Even though most of the research discoveries have not been implemented in actual clinical practice, compiled insights have been drafted with a deepened understanding of the genetic components, their effects, and interactions, as well as other factors such as environmental factors. In the past decades, ALS-associated genes have been discovered, and disease biomarkers such as neurofilaments have demonstrated great clinical relevance with supporting clinical trials[21,77]. Due to the overall heterogeneous causes and disease outcomes, and the short survival time after disease onset[21] with wide survival variations, there is still much more effort that is needed to be done. In terms of genetic factors, utilizing the current advancements in biotechnologies and also combining different omics to understand the complex interplays between different molecular layers (using complexity to fight complexity), would greatly speed up the research process in the genetic components contributing to understanding the disease.

3 RESEARCH AIMS

The overall aim of this thesis is to discover novel omics signatures of complex diseases by analyzing various types of high-throughput high-dimensional molecular data. In detail, the thesis aims to:

1. Construct a single cell-level transcript quantification tool;
2. Discover existing or novel irQTL across different brain regions;
3. Determine phenotypic differences between ALS patients and controls to identify unique disease signatures through multi-omics data analysis.

4 MATERIALS AND METHODS

4.1 MATERIALS

4.1.1 ScRNA-Seq Data

Study I focused on the development of a transcript quantification method in a form of a stand-alone Linux-based tool for single-cell transcriptomics data. To measure quantification accuracy and benchmark real datasets, simulated scRNA-Seq data was used for accuracy assessment based on true counts followed by further validations based on real scRNA-Seq data. For all datasets, mapping was done based on the human reference genome hg38.

4.1.1.1 Simulated Dataset

We simulated a scRNA-Seq dataset to mimic the output from 10X Genomics 3' sequencing (v3.1). Polyster[79] v1.24.0 was used to generate fastq reads with a positional bias model to produce reads with highly skewed read distribution near the 3' end. The mean fragment length and its standard deviation were also modeled and fitted into Polyester before reads generation. To infer transcript-count ratios at the single-cell level with better accuracy, we made observations for each gene using a full-length Smart-Seq2 dataset[11] to avoid any quantification bias introduced by prime-sequencing. Sequencing depth was modeled using a PBMC sample from a healthy human donor, Single Cell Gene Expression Dataset using Single Cell 3' v3, Chromium Connect Channel 1, 10x Genomics (2020, February 28). The above procedure accounted for the simulation of paired-end fragments. For 10X Genomics, read 1 is mainly made up of cell barcode sequence and unique molecular identifier (UMI) sequence for each RNA molecule of each cell, and read 2 contains the actual RNA sequences. Therefore, read 1 generated from Polyester was further modified to a fixed length containing both the barcode and UMI sequences. Barcodes are selected at random for inclusion based on the barcode list from 10X Genomics, which contains around 3 million unique barcode sequences iteratively used by the technology for their experiments under the same protocol version (v3.1) and are 16 base pairs (bps) in length. The UMI sequences are usually 12bps long for the same version (v3.1) and were modeled using random nucleotide sequences. In a real sequencing setting, PCR errors will be observed and corrected, and therefore we did not consider such modeling. In total, 3,995 cells were produced for the simulated dataset.

4.1.1.2 Real Dataset

We utilized a bone-marrow CITE-Seq scRNA-Seq sample from Stuart *et al.*[80] for benchmarking our method against other quantification methods. To ensure fair comparisons, the same number of cells (n = 20,840 cells) containing matching barcodes across these comparing methods were used. Homogenous cell type annotations were applied to all post-

quantification outputs. To validate our findings, 15 human fetal bone-marrow samples containing 3,055 cells[81] from Smart-Seq2 were used.

4.1.2 Dataset from the GTEx Project

For Study II, raw RNA-Seq samples (n = 1,191) from 172 individuals across 13 human brain regions were obtained from the GTEx project (dbGaP Accession phs000424.v7.p2.c1)[82]. Samples were collected from disease-free sites and individuals with infectious diseases such as the human immunodeficiency virus (HIV) infection and metastatic cancer, or therapeutically treated with chemotherapy or radiotherapy were exempted[82]. The age group of the donors ranged from 21 to 70 years. Apart from the RNA-Seq samples, whole-genome sequencing (WGS, phs000424.v7.p2.c1) data from these donors were also included in our study. In total, the WGS dataset consists of 5,987,177 105 single-nucleotide polymorphisms (SNPs) and 509,531 insertions and deletions (Indels). We compared systematically our irQTL findings with the brain eQTL (phs000424.v7.p2.c1) and sQTL (phs000424.v8.p2) results from the GTEx project.

4.1.3 Human Genotype-Phenotype Associations from PhenoScannerV2

We used publicly available PhenoScanner V2, a database containing over 65 billion genotype-phenotype association results, including diseases-trait associations of mQTL, eQTL, pQTL, and DNA methQTL[83]. Association results for eQTL were used to find relevant associations of irQTL with phenotypic traits.

4.1.4 ALSrisc Cohort

Study III is based on the ALSrisc study which stands for Biomarkers and Risk Factors for Amyotrophic Lateral Sclerosis. The ALSrisc cohort includes all newly diagnosed ALS patients (cases) from 2016 onwards in Stockholm county, as well as two control groups, namely disease-free full siblings and spouses of the ALS patients. ALS patients are recruited at the ALS Clinical Research Centre of Karolinska University Hospital, which is the only tertiary center for ALS in Stockholm. All patients received a diagnosis of probable, possible, or definite ALS, according to the El Escorial criteria. ALS patients, sibling controls, and spouse controls are all enrolled at the

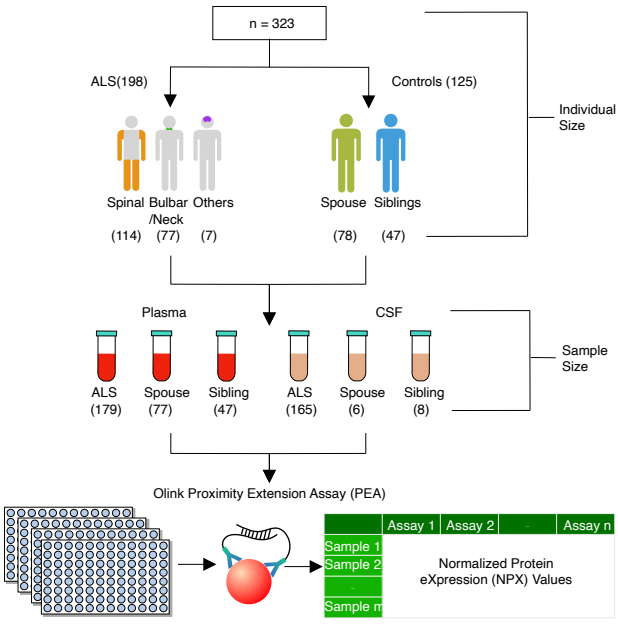

Figure 1. Study participants, sample preparation, and experimental procedures for Study III. Adapted from Lu P. *et al.*, manuscript in preparation[2].

time of ALS diagnosis or shortly thereafter. For Study III, we included a total of 198 ALS patients, 78 sibling controls, and 47 spouse controls (**Figure 1**).

4.1.4.1 ALS Plasma and CSF Proteomics

Study III aims to discover protein biomarkers associated with ALS diagnosis, prognosis, functional declination, etc., based on the assessment of CSF and plasma samples of the ALS patients and normal controls from the ALSrisc cohort. For ALS patients, 179 plasma samples and 165 CSF samples were collected. For the controls, 77 plasma and 6 CSF samples were acquired for the spouse controls, and 47 plasma, as well as 8 CSF samples, were obtained from the sibling controls (**Figure 1**). All samples have undergone proteomics profiling with the Olink proteomics technology (www.olink.com).

4.2 EXPERIMENTS

4.2.1 Sample Collection and Processing

Blood and CSF biopsies were collected from all ALS patients either at the time of diagnosis or within three months thereafter, and annually thereafter. The collection was carried out via lumbar puncture. Blood samples from the sibling and spouse controls were collected at the time of diagnosis of their corresponding ALS sibling or partner. CSF samples were also collected among the controls recruited in the pilot phase of the ALSrisc study in 2015 during their blood sample collections. To isolate plasma, blood samples were centrifuged at room temperature at 2000g for 10 min. CSF samples were centrifuged at 400g for 10 min at 4°C. Samples were stored as approximal 800µl - 1000µl aliquots at -80°C directly after collection. Both blood and CSF samples were thawed one time to prepare 70µl aliquots each for further sequencing using the Olink Proteomics platform.

4.2.2 Genotyping and C9 Typing

Genotyping was performed using Illumina Infinium Global Screening Array (GSAMD-24v3-0-EA_20034606_A1) on the PLUS strand of DNA, which characterizes approximately 730,000 SNPs, using the SNP&SEQ Technology Platform in Uppsala, Sweden. Sanger Imputation Service was used for genotype imputation. Quality control filters were applied before and after the imputation. Hexanucleotide repeat expansion (G4C2) mutation in intron 1 of the C9ORF72 gene is associated with familial ALS. We identified homozygote carriers using fluorescent PCR as previously described by DeJesus-Hernandez *et al.*[84]. In the case of homozygote carriers, we used triplet primer PCR to quantify the expansion of the repeats.

4.2.3 Olink Proximity Extension Assay (PEA)

We profiled over 300 unique proteins for both the CSF and plasma samples using the proximity extension assay technology, a high-throughput immunoassay from the Olink Proteomics AB, Uppsala, Sweden (www.olink.com). The proteins were sequenced across four panels, namely, 1) Olink Target 96 Inflammation, 2) Olink Target 96 Neurology, 3) Olink Neuro Exploratory, and 4) Olink Target 96 Cardiovascular III panels, with each channeling 92 proteins per panel.

4.2.4 Study I

We utilized mainly the concept of equivalence class[85,86] and the use of alternating expectation-maximization (AEM) algorithm[87,88] in Study I. Equivalence classes[85,86] (eqClass) are clusters of highly-similar transcripts and reads mapped similarly to a set of transcripts within each cluster constituted an aggregated read count known as the transcript compatibility count (TCC)[85,86]. The TCC of each eqClass served as a basis for the tabulation of transcript abundance using the AEM algorithm.

4.2.4.1 Modeling Single-Cell Transcript Counts using the AEM Algorithm

Quantifying transcript abundance in Study I revolves around the estimation of transcript abundance using the AEM algorithm[88]. For each cell, we constructed a vector y, summarizing read counts r_i of TCCs of all the eqClasses of the cell. The vector y can be derived from the read alignment of raw fastq files using Alevin[89], and could alternatively be done using other methods such as Kallisto-bustools[90] as these tools produce read counts in the form of eqClasses, which can be served as inputs to the quantification of isoforms in our algorithm (**Figure 2**). The presumption here is that y follows a Poisson distribution with mean μ, in which, vector μ is the expected read counts mapped to all eqClasses in a cell, and μ follows the bilinear model,

$$\mu = X\beta,$$

such that β models the expression counts for each isoform and X matrix summarizes the read sharing between the isoforms in each eqClass. To optimize algorithmic efficiency, transcripts from the same eqClass were grouped under the same transcript cluster (TC) and the estimation of transcript abundance was done independently for each TC. As AEM is an iterative algorithm that alternates and updates the values of X and β between the expectation (E) and maximization (M) steps, given Y, a starting matrix for X is needed to compute the first set of β values. The initial guess of X matrix was constructed by first, constructing a simulated set of scRNA-Seq data, modeling the prime-bias problem introduced by the droplet-based method (here we focused on 10X Single Cell 3' protocol) based on the same criteria for generating the simulated data mentioned in *Section 4.1.1.1*. Similarly, this was followed by the mapping procedure and transcripts from the same eqClass were grouped under the same TC and will be normalized and used as an initial X matrix independently for each TC in the AEM algorithm (**Figure 2**). For the X matrix, a threshold was set to merge isoforms with highly similar sequences that are statistically unidentifiable from each other (here we called them isoform paralogs). Given X and Y, for each cell and each TC, the E step estimates the transcript abundance β. Once β is estimated with an initial set of transcript abundance values, given β and Y, the values in the X matrix get updated with a new set of TCC values in the M step. The updates to the values of β and Y reiterates between E and M steps until β converges.

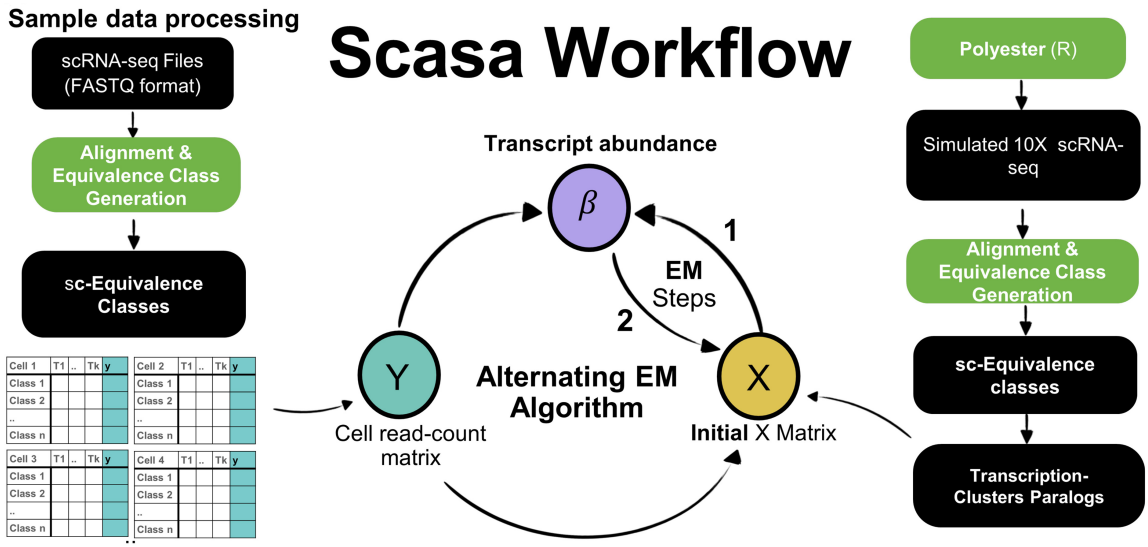

Figure 2. Algorithmic workflow for Scasa, the tool developed in Study I to estimate transcript abundance at the single-cell level. Adapted from Lu P. *et al.*, 2021[91].

4.2.5 Study II

4.2.5.1 Isoform Quantification and Ratio Estimation

We acquired demultiplexed RNA-Seq fastq files of the brain samples from the GTEx project[60,82] and carried out mapping and isoform quantification using XAEM v0.1.0[88] and human genome reference hg19. To identify genetic regulation of isoform expression, we considered only isoforms from multi-isoform genes, which retained a total of 31,482 isoforms from 9,401 genes out of a total of 46,719 isoforms from 24,629 genes. Out of which, around 90% of the isoforms are from the protein-coding genes, and the rest include long non-coding RNAs (lncRNAs), anti-sense RNAs, etc. To reduce redundancy, samples with half or more than half of the isoforms having zero counts were removed. Raw isoform counts were normalized to transcript per million (TPM) counts and isoform ratios were estimated for each isoform, by dividing the TPM of each isoform by their respective gene-level TPM. These TPM isoform ratios served as the phenotypes for the estimation of irQTL in GWAS analysis in the next step.

4.2.5.2 irQTL Discovery

Before the GWAS analysis, we examined the genomic kinship using PLINK v1.9[92] and used the first three principal components (PCs) as the covariates to correct the isoform ratios using linear regression. Apart from the PCs, age, and sex were also considered in the regression model. Using the covariates-corrected isoform ratios as phenotypes, we performed the GWAS analysis on their corresponding genotype data using RegScan v0.5[93] which enables fast analysis for large datasets. We considered only *cis*-regulatory loci and each *cis*-region locus was defined as ± 1Mb up- and down-stream around the corresponding gene. SNPs with the lowest p within each locus were selected as the lead variant and associations having $p < 5 \times 10^{-8}$ were retained for subsequent analyses. To identify splicing regulation-specific QTL, we compared

the association results with the *cis*-eQTL from the GTEx project[60,82], and retained irQTL with eQTL $p > 0.05$ as the final set of irQTL[1].

4.2.5.3 Stratified linkage disequilibrium score regression (S-LDSC)

We utilized S-LDSC v1.0.0 to examine if the annotated genetic regions are enriched for heritability of a certain trait based on the GWAS summary statistics from the LG-Hub[94]. Harmonizing of the summary statistics was carried out using the same software. The heritability enrichment score was defined as the proportion of heritability captured divided by the proportion of annotated SNPs. To increase the sensitivity of our estimation, we controlled for residual variance recommended by LDSC[95,96], to fit the LDSC-v1.2 baseline annotations as covariates. The whole process was done separately for each brain tissue to avoid any possible multi-collinearity introduced by the similarities between the brain tissues.

4.2.5.4 Mendelian Randomization (MR) analysis

Neuro-related traits from the LD-Hub[94] and GWAS results of the UK Biobank (UKB) diseases from Neale's lab[38] were retrieved and duplicated traits as well as highly correlated traits after comparison between the data sources, were removed. MR analysis was carried out between the normalized isoform ratios and the phenotypic traits using the standard inverse-variance weighted (IVW) method for all *cis*-irQTL.

4.2.6 Study III

Demographically, we compared the sex, age, BMI, smoking, hypertension, etc. of ALS patients to the controls, as well as the clinical characteristics of study participants to all ALS patients in Stockholm during the study period using the Motor Neuron Disease (MND) Quality Registry[97] to assess study population representativeness.

4.2.6.1 Proteins between cases and controls

We utilized the generalized estimating equation (GEE) model[98] to compare protein concentrations between ALS patients and controls while accounting for possible clustering effects between individuals from the same family. To avoid potential false positives due to multiple testing, we applied the Benjamin-Hochberg (BH) False-Discovery Rate (FDR) to adjust p-values. Differential expression was determined if a protein showed a concentration difference of normalized protein expression (NPX) > 0.25 between cases and controls at FDR < 0.05. We analyzed plasma and CSF separately and compared proteins in plasma between ALS patients and their siblings as well as spouse controls. Correlations between plasma-based and CSF-based concentration differences of each protein were calculated to understand potentially different behaviors of proteins in the periphery and intracranial compartment. Multivariable adjustments were made, including age, sex, technical factors, and factors related to both the risk of ALS and protein concentrations. Missing values in BMI, smoking, and hypertension were imputed before all analyses using simple or multiple imputation methods.

4.2.6.2 Proteins and risk of death and survival among ALS patients

We aimed to identify potential protein biomarkers in plasma or CSF samples that could serve as prognostic indicators of disease in ALS patients. To achieve this, we performed several analyses on ALS patients, all of whom were followed up through the MND Quality Registry from the time of diagnosis until death or January 31st, 2022, whichever occurred first. During this period, 161 out of 198 ALS patients died, with a median survival of 506 days from diagnosis (911 days from symptom onset). We used a Cox proportional hazards model to assess the risk of death per NPX increase of each protein in plasma and CSF, both with unadjusted and multivariable models. The multivariable model was adjusted for covariates such as age, sex, number of freezing days, sequencing plate identity, body mass index (BMI), smoking, and hypertension, as well as other known prognostic indicators for ALS, including the site of onset, revised ALS functional rating scale ALSFRS-R[99] score at diagnosis, and diagnostic delay (the time interval between symptom onset and diagnosis). In the secondary analysis, we categorized ALS patients into high or low levels of specific proteins and used maximally selected rank statistics[100] to determine the optimal NPX cut-off value for each protein. This allowed us to determine the best separation between the two groups of ALS patients in terms of survival. Kaplan-Meier survival curves were then plotted for patients with high or low levels of each protein, and the difference between the survival curves of the two groups was assessed using a log-rank test to determine its statistical significance.

We conducted a secondary analysis to determine the summary effect of proteins that were significantly associated with the risk of death following an ALS diagnosis. We computed a poly-protein risk score (PRS) by summing the NPX values of each protein, weighted by their estimated $\hat{\beta}$ coefficients derived from the corresponding Cox model. Mathematically, this PRS can be modeled as:

$$PRS_{ij} = \sum_{k}^{N} \hat{\beta}_{jk} x_{jk}$$

where $\hat{\beta}_{jk}$ is the estimated coefficient of Cox model for a protein k in setting i for sample j whereas x is the NPX value of protein k for sample j. We only included proteins that were significantly associated with the risk of death in the multivariable Cox models, and we calculated PRS values in plasma and CSF separately. ALS patients were then classified using maximally selected rank statistics as having either high or low PRS, and were done independently for plasma and CSF samples.

4.2.6.3 Proteins and functional decline among ALS patients

Apart from analyzing the risk of death and survival, we also investigated the relationship between proteins and the rate of functional decline, as measured by the ALSFRS-R score against time since diagnosis. The analysis was conducted separately for plasma and CSF. For each protein, ALS patients were divided into high or low-protein-level groups based on their NPX values using maximally selected rank statistics as previously described. Differences in

the rate of ALSFRS-R decline between the two patient groups were assessed using a generalized additive model. ALSFRS-R scores were treated as the response variable, and the patient group was treated as the predictor variable, with thin plate smooth splines. To account for individual patient differences in the analysis, patient identity was modeled as a random effect, with an interaction term between patient identity and time.

4.2.6.4 Longitudinal protein profiles among ALS patients

We also investigated the longitudinal profiles of proteins in plasma and CSF after ALS diagnosis. Linear mixed-effects models were used, with NPX values of each protein as the response variable, and time after diagnosis as the predictor variable, adjusted for the same set of covariates mentioned in the previous analyses. Sample identity was included as a random effect in the model. FDR correction was applied to adjust for multiple testing, and a significance level of FDR < 0.15 was used due to the smaller sample size as compared to the analysis of ALS cases and controls. Yet, a final threshold of FDR < 0.05 was still used to filter for significance owing to the larger number of proteins showing statistically significant trends over time.

4.3 ETHICAL CONSIDERATIONS

Study I used published datasets, as well as simulated datasets, and therefore there was no handling of sensitive data. For the RNA-Seq and WGS datasets from the GTEx project used in Study II, permission was granted to access the data and no sensitive data leading to the identification of individuals was used. Study III was approved by the Ethical Review Board in Stockholm, Sweden (DNRs 2014/1815-31/4 and 2018-1065/31). Oral and written informed consent was granted from all study participants.

5 RESULTS

5.1 STUDY I

We developed a tool named Scasa[91], with the method to estimate isoform abundance for single-cell droplet-based transcriptomics technologies. For this study, we addressed the prime-bias problem introduced by these droplet-based protocols and benchmarked Scasa quantification results with other existing isoform and gene quantification software. Scasa outperforms other methods based on simulated data and quantifying isoform expression for a bone marrow dataset[80] revealed a novel CD14 monocyte subtype missed by gene expression quantifications.

5.1.1 Addressing the Prime-Bias Problem Introduced by Droplet-Based Technologies

Based on the hg38 transcriptome reference, 52,046 (73.3%) of the isoforms are separately quantifiable (such that they do not form isoform paralogs, and are of paralogs size = 1) for bulk RNA-Seq data, and the remaining contains paralogs. The prime bias problem in the scRNA-Seq data reduced the number of separately quantifiable isoforms to 21,287 (30.0%)[91]. The cost to such a problematic event is in turn the high similarities in their X matrix distributions, and created estimation problem as they are statistically indistinguishable. Not accounting for these paralogs could result in bias estimates in addition to the prime-bias phenomenon, which violates the read-sharing symmetry assumption presumed by bulk RNA-Seq methods. Scasa identified and quantified the paralogs present in the data and can provide more accurate quantification by modeling the prime-bias effect.

5.1.2 Scasa Outperforms Other Quantification Methods

We benchmarked Scasa v1.0.0[91], using the simulated dataset, against both the isoform-level (IL) and gene-level (GL) quantification methods to evaluate its quantification performance. These included (i) single-cell GL quantification tools, Kallisto-bustools[90,101] (Kallisto v0.46.1 and bustools v0.39.3), Alevin (v1.4.0[89]), Cellranger (v3.1.0[102]), STARsolo(v2.7.2b[16]); (ii) bulk IL and GL quantification tools, Kallisto v0.46.1[101], Salmon v1.4.0[103]; and (iii) bulk IL quantification tool, Terminus v0.1.0[104]. To compute comparative GL values for the IL methods, we summed the component isoforms based on human genome reference hg38. Scasa has demonstrated its superior performance in both isoform and gene expression quantification settings, with Salmon being the closest competitor at the IL level. Among the gene-level quantification methods, Alevin and STARsolo also performed well compared to Scasa (**Figure 2**).

5.1.3 Novel CD14 Monocyte Subtype Revealed by Scasa

We used publicly-available bone marrow mononuclear dataset[80] with measured antibodies with well-defined cell-type identification and annotations. At IL quantification, Scasa (isoform) detected a distinct CD14 monocyte subgroup which was not discoverable by gene-level quantification using Scasa-gene, Cellranger, and Alevin (**Figure 2A**). Notably, among the statistically significant differentially-expressed (DE) isoforms of this monocyte population, the

top four DE isoforms came from the *TYROBP* gene[91] and they displayed unique isoform expression signatures within the specific population of cells. The phenomenon was diluted at the GL expression quantification (**Figure 2B**).

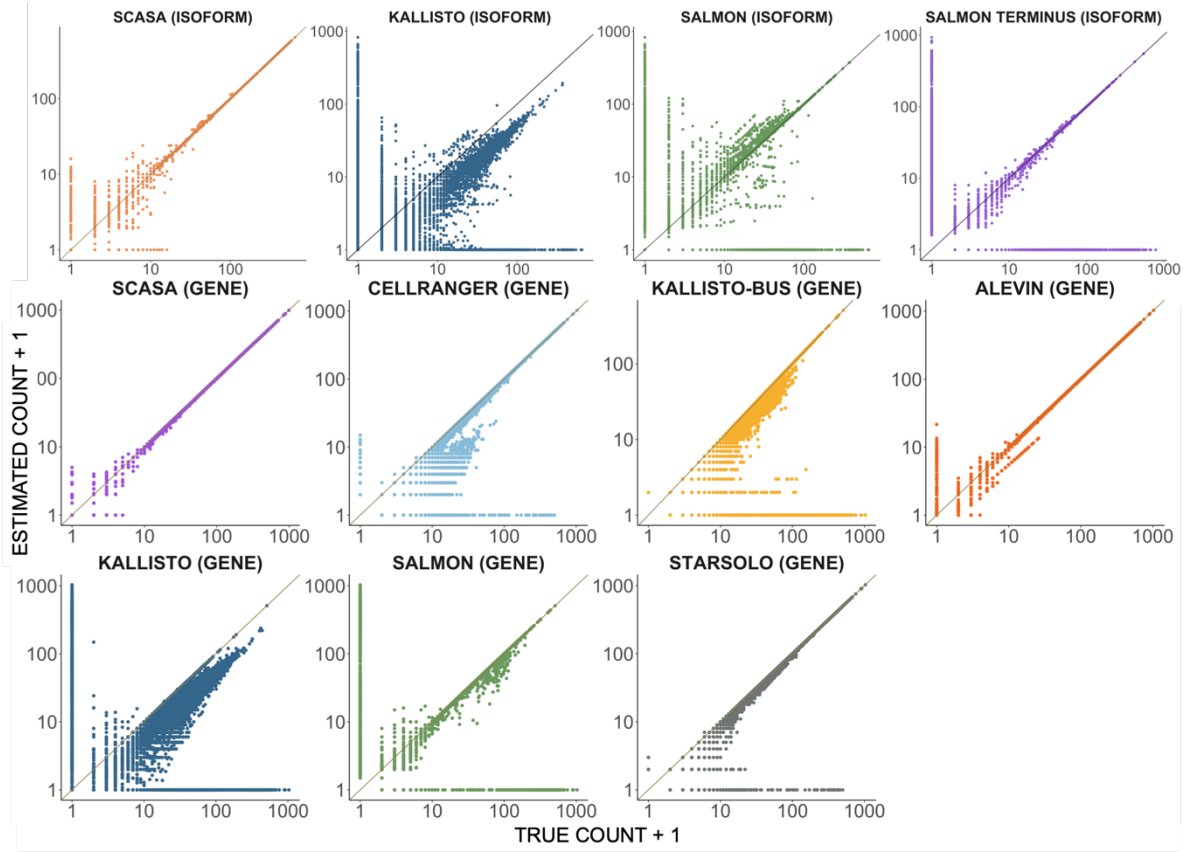

Figure 2. Comparison of Scasa against competing methods for both IL and GL quantifications. Simulated data with true counts were used for the comparison, where the x-axis represents the true counts and the y-axis denotes the estimated counts by the software. Adapted from Lu P. *et al.*, 2021[91].

5.2 STUDY II

5.2.1 *Cis*-irQTL Discovery

Study II identified splicing-specific 4,241 *cis*-irQTL across the 13 brain regions, as compared to the eQTL gene-level result from GTEx ($p_{irQTL} < 5 \times 10^{-8}$ and $p_{eQTL} > 0.05$). We cross-referenced the *cis*-irQTL with the sQTL discovered by sQTLseekeR[105] from the GTEx project and overlapping of 1,126 QTL were found in sQTL a$p < 5 \times 10^{-8}$ significance. We have also compared *cis*-irQTL with the THISTLE[106] sQTL results, in which THISTLE employ a gene-wise heterogeneity test for sQTL discovery, i.e., splicing genes (sGenes). A total of 96 sGenes were found to be associated with the *cis*-irQTL sGenes (n = 874).

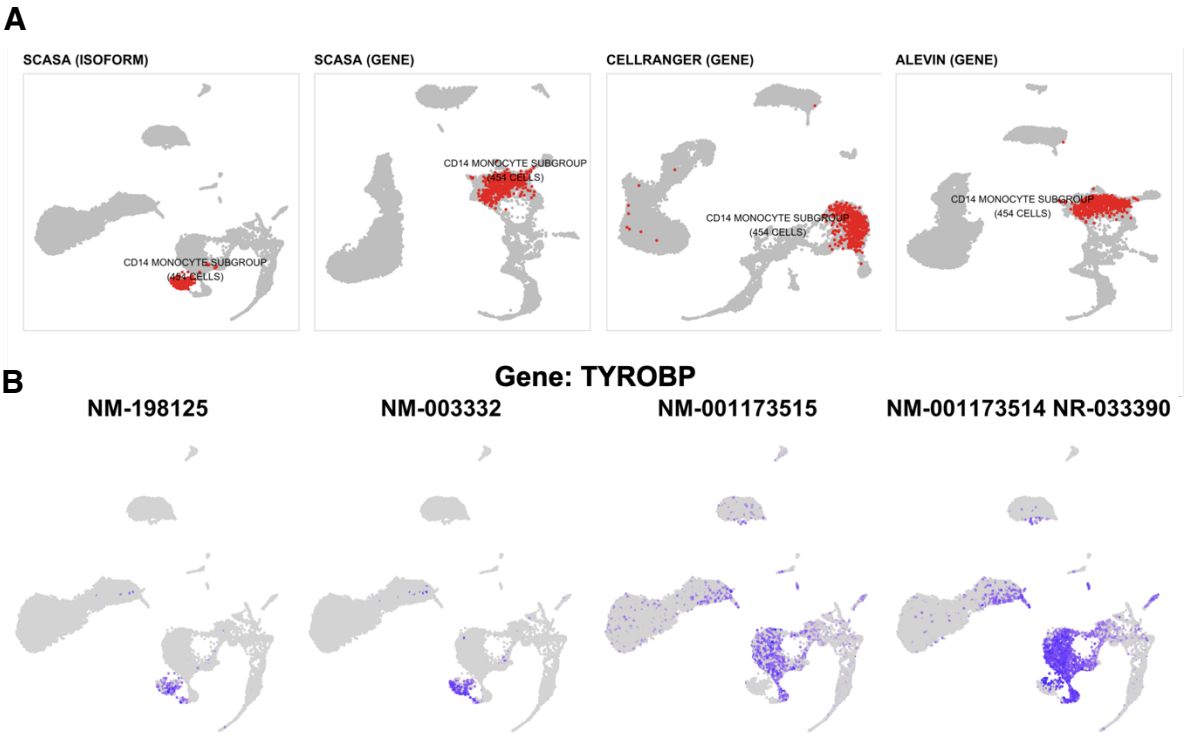

Figure 3. The novel CD14 monocyte subtype revealed by Scasa. (**A**) Cells highlighted in red are the cells from the new CD14 monocyte population discovered by Scasa. These cells diffused back into the main CD14 monocyte cluster at GL quantification. (**B**) Isoform expression levels for one of the top DE-isoform-associated genes, TYROBP. A distinct expression pattern was observed, within the new CD14 monocyte sub-cluster. Adapted from Lu P. *et al.*, 2021[91].

In terms of the *cis*-irQTL, similar to the *cis*-eQTL results from GTEx, most of them are found near the transcription start sites (TSS) and exhibited a wider distribution around TSS as compared to that of the *cis*-eQTL (**Figure 3**). The difference could be a complication of differential splicing patterns across isoforms from the same gene as well as lower statistical power compared to gene expression QTL due to the ambiguity arising from read-sharing between isoforms.

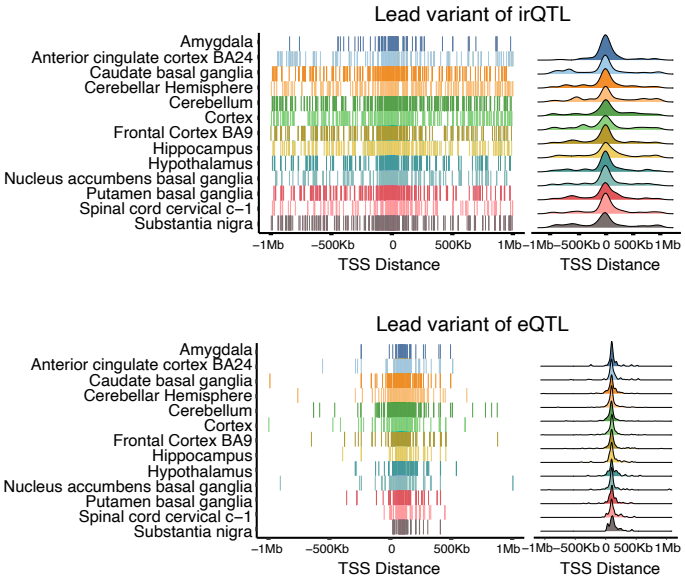

Figure 3. Distribution of the *cis*-irQTL and *cis*-eQTL around TSS regions across 13 brain tissues. Adapted from Lu P. et al. (in press)[1].

5.2.2 Heritability Enrichment and MR Analyses

A total of 1,482 enrichment tests were carried out to test for heritability enrichment of phenotypic traits across the 13 brain tissues. In terms of neuro-related, three brain tissues were found to possess significant associations, including Alzheimer's or dementia, mood swings, nervous feelings, sensitivity or hurt feelings, sleep duration, alcohol intake, and contraceptive pill intake with the frontal cortex (BA9); educational attainment, alcohol intake, intelligence, and knee pain with the cortex; and anxiety or depression with the cervical spinal cord.

The MR analysis found 250 isoform-disease pairs with causal relationships at FDR < 0.05. Traits enriched include educational attainment, sleep, psychiatric disorders, feelings, and alcohol intake. Gene *MMAB*, with the most MR discoveries, includes traits such as sleep duration, insomnia, neuroticism, miserableness, and schizophrenia. This was in line with the literature as Vitamin B12 is involved in the production of sleep-regulating neurotransmitter melatonin[107,108], and the protein product of *MMAB* was reported to catalyze the conversion of vitamin B12 into its final product adenosylcobalamin[109].

5.3 STUDY III

5.3.1 Demographics and the Representativeness of the ALSrisc Cohort

Comparing the ALS patients and the controls, the ALS patients were on average, slightly older in age, had lower BMI, and showed a higher prevalence of C9orf72 mutation compared to the controls (**Table 1**). Comparing the ALSrisc cohort to the Stockholm population, the characteristics of the ALS patients in the present study are comparable and representative of the entire ALS population in Stockholm during the study period in terms of site of onset, ALSFRS-R score, diagnostic delay, and mutation status (**Table 2**).

5.3.2 Protein Biomarkers Discovery

For all comparisons, plasma and CSF samples were assessed independently. We discovered (i) potential protein biomarkers that differ significantly between ALS cases and controls; (ii) biomarkers related to the risk of death and survival among ALS patients; (iii) biomarkers with significant association with ALS functional decline; and (iv) longitudinal proteins that changed significant overtime after diagnosis. Among the top proteins, neurofilament light chain (*NEFL*), showed the strongest association across (i), (ii), and (iii) in both plasma and CSF samples. Longitudinally, however, *NEFL* did not demonstrate highly varying expression levels after disease diagnosis. Other than *NEFL*, varying top protein biomarkers were observed, Higher plasma protein levels such as *EDA2R*, *TNFRSF12A,* and *MB* as well as higher CSF levels of *GDF-15* were associated with a higher risk of ALS, whereas higher plasma levels of *RGMA*, *GDF-8,* and *MMP-3* were associated with a lower risk of ALS. Higher plasma levels of *EDA2R*, *TNFRSF12A,* and *GDF-15* were associated with a faster functional decline whereas

higher plasma levels of *ITGB2* and *GCP5* were associated with a slower functional decline, after ALS diagnosis. Higher CSF levels of *MB* were associated with a higher risk of death. After the ALS diagnosis, *EDA2R* and *GDF-15* showed an increment in concentration whereas *RGMA, GDF-8, MMP-3, ITGB2,* and *GCP5* showed a decreasing concentration in plasma over time since ALS diagnosis.

Characteristics	ALS patients (M = 198)		Spouse controls (M = 78)		Sibling controls (M = 47)	
	Plasma (N = 179)	CSF (N = 165)	Plasma (N = 77)	CSF (N = 6)	Plasma (N = 47)	CSF (N = 8)
Sex, N (%,1d.p.§)						
Female	89 (49.7%)	79 (47.9%)	46 (59.7%)	4 (66.7%)	25 (53.2%)	2 (25.0%)
Male	90 (50.3%)	86 (52.1%)	31 (40.3%)	2 (33.3%)	22 (46.8%)	6 (75.0%)
Age (mean±SD, 1d.p.), year	65.2±11.1	65.1±11.4	63.9±10.1	55.0±10.9	61.2±8.3	59.8±7.6
BMI (mean±SD, 1d.p.)*	23.9±3.9	23.9±4.0	25.7±3.6	25.4±2.3	25.4±4.4	25.7±3.6
Hypertension, N (%,1d.p.)						
No	80 (44.7%)	71 (43.0%)	47 (61.0%)	3 (50.0%)	31 (66.0%)	4 (50.0%)
Yes	60 (33.5%)	47 (28.5%)	20 (26.0%)	3 (50.0%)	15 (32.0%)	4 (50.0%)
Unknown	39 (21.8%)	47 (28.5%)	10 (13.0%)	0 (0.0%)	1 (2.1%)	0 (0.0%)
Smoking, N (%,1d.p.)						
Never smoker	61 (34.1%)	54 (32.7%)	26 (33.8%)	4 (66.7%)	23 (48.9%)	3 (37.5%)
Former smoker	70 (39.1%)	60 (36.4%)	38 (49.4%)	2 (33.3%)	19 (40.4%)	4 (50.0%)
Current smoker	12 (6.7%)	7 (4.2%)	5 (6.5%)	0 (0.0%)	4 (8.5%)	1 (12.5%)
Unknown	36 (20.1%)	44 (26.7%)	8 (10.4%)	0 (0.0%)	1 (2.1%)	0 (0.0%)
Sample freezing days (mean±SD, 0d.p.)	979±406	1001±385	976±395	1267±248	986±367	1101±301

Table 1. Baseline characteristics of patients with ALS and their spouse and sibling controls. M represents the number of individuals and N represents the number of samples. §Numbers are rounded off to 1 decimal place (d.p.). *BMI is expressed in the unit of kg/m2. Imputed values for missing data are included for BMI, hypertension, and smoking. Adapted from Lu P. *et al.*, manuscript in preparation[2].

Characteristics	ALS patients in study (M = 198)		Stockholm ALS population (M = 504)
	Plasma (M = 179)	CSF (M = 165)	
Onset type, N (%,1d.p.§)			
Spinal	107 (59.8%)	95 (57.6%)	304 (60.3%)
Bulbar / Neck	66 (36.9%)	65 (39.4%)	149 (29.6%)
Other	6 (3.4%)	5 (3.0%)	30 (6.0%)
Unknown	0 (0.0%)	0 (0.0%)	21 (4.2%)
Disease progression at diagnosis, N (%,1d.p.)*			
Slow	29 (16.2%)	22 (13.3%)	93 (18.5%)
Intermediate	69 (38.5%)	56 (33.9%)	97 (19.2%)
Fast	80 (44.7%)	83 (50.3%)	136 (27.0%)
Unknown	1 (0.6%)	4 (2.4%)	178 (35.3%)
ALSFRS-R score at diagnosis (mean±SD,1d.p.)	37.2±7.5	37.8±7.4	34.2±10.4
Diagnostic delay (mean±SD,1d.p.), months	16.8±16.4	15.6±15.6	18.0±19.8
Gene mutation status, N (%,1d.p.)+			
SOD1	3 (1.7%)	3 (1.8%)	3 (0.6%)
C9ORF72	20 (11.2%)	16 (9.7%)	21 (4.2%)
OPTN	1 (0.6%)	1 (0.6%)	1 (0.2%)
ATXN8	1 (0.6%)	1 (0.6%)	1 (0.2%)

Table 2. Comparison of characteristics between ALS patients included in the study and the entire ALS population in Stockholm. M represents the number of ALS patients. §Numbers are rounded off to 1 decimal place (d.p.). *Slow progression means a decline of <0.5 ALSFRS-R score per month, intermediate progression means a decline of 0.5 to 1.1 ALSFRS-R score per month, and fast progression means a decline of >1.1 ALSFRS-R score per month. +Number of patients tested for mutation status of the listed genes is 161 in the present study and 197 in the entire ALS population in Stockholm. Adapted from Lu P. *et al.*, manuscript in preparation[2].

6 DISCUSSION AND CONCLUSION

6.1 STUDY I

The purpose of alternative splicing is to generate isoforms with alternative biological functions. Bulk RNA-Seq methods are not effective for quantifying isoforms in scRNA-Seq datasets due to the 3' bias in the droplet-based sequencing protocols. We developed Scasa to accurately quantify isoform expression in 10X 3' scRNA-Seq data by addressing this bias while utilizing the concept of transcription clusters and isoform paralogs. Scasa outperformed other popular methods and identified isoform-specific cellular subsets in a case study of bone marrow cells. Scasa has the potential to identify biomarkers and cellular subsets not detectable at the gene level.

Even though Scasa is constructed based on the Chromium Single Cell 3' 10X Genomics technology, the tool can be extended to work on other technologies which also use prime tagging techniques. In addition, Scasa can be easily used as the downstream quantification tool for gene-level post-alignment outputs from Alevin and Kallisto-bustools. On the other hand, due to shared high similarities in certain isoforms, not all isoforms can be estimated individually and were grouped into isoform paralogs by Scasa. This however in term limits the number of identifiable isoforms. Characterizing the isoform members of a paralog requires additional information and possibly experimental validation, which is beyond the scope of our study.

6.2 STUDY II

In Study II, we used GWAS analysis on isoform ratio normalized TPMs to identify irQTL, which could not be detected by analyzing gene-level expressions alone. By studying the genetic architecture of irQTL, we discovered that isoform ratios are involved in regulating educational attainment in multiple tissues, including the frontal cortex, cortex, cervical spinal cord, and hippocampus, which are also associated with various neuro-related traits and diseases. Our MR analysis identified 1,139 pairs of isoforms and neuro-related traits with plausible causal relationships, indicating that investigating overall gene expressions could miss critical transcript-level biomarkers in the human brain for neuro-related complex traits and diseases. The study focused on brain tissue due to its abundant alternative splicing events and specific functions linked to brain-related phenotypes. While assessing more tissue samples from the GTEx project could be valuable, it would require significant computational resources. Future studies could incorporate more tissues into investigating and discovering other tissue-specific irQTL. Additionally, since whole-blood RNA sequencing data are available in multiple human cohorts, we foresee a consortium-based investigation of irQTL associated with various human complex traits and diseases.

We discovered brain *cis*-irQTL that controls isoform proportions which is different from eQTL. Genes of these irQTL are enriched for heritability in neuro-related traits. Our analysis demonstrated the importance of quantifying isoform expressions, as some genetically regulated

functional transcripts may only be detected by utilizing the splicing information, and have downstream effects on phenotypes via MR analysis. We utilized XAEM for isoform quantification which allows for accurate quantification of isoform expression for multi-isoform genes[88]. The genome references such as the Ensembl[110] and the GENCODE[111], contain much more non-curated isoforms, and many of the isoforms from the same gene only by a few bases in sequencing, accentuating the difficulties in isoform quantification. Therefore, using curated isoforms from RefSeq is more suitable for the study.

The discovery of irQTL and its associated traits indicated that it is important to explore the effect introduced by alternative splicing rather than looking at the overall gene expression phenotypes. Besides, it is important to differentiate between sQTL and irQTL. sQTL is investigating the direct alternative splicing events while isoform expression itself can be treated as splicing phenotypes. Studying isoform expressions as phenotypes could provide direct downstream information on sQTL.

6.3 STUDY II

Our results on *NEFL* indicated that *NEFL* is a notable biomarker for disease risk, functional decline, and risk of death after ALS diagnosis in both plasma and CSF. The comparable findings between the two sample types support the notion that NEFL measured in the periphery can serve as a reliable surrogate for the intracranial compartment[112]. However, *NEFL* is not a highly specific biomarker for ALS, as it has also been observed in other neurodegenerative diseases[113-116]. *NEFL* does not show clear temporal trends over time since ALS diagnosis, in agreement with existing literature[112,117,118], limiting its use as a biomarker for clinical trials.

The study discovered *EDA2R*, *TNFRSF12*, *MB*, and *GDF-15* as *potential risk proteins* that higher plasma levels of these proteins are associated with a higher risk of ALS and functional decline. *EDA2R* has been shown to increase with age[119,120] and is linked to loss of neuronal functions. Its upregulation in ALS suggests a role in muscle cell survival and catabolism. *EDA2R* may be a biomarker for muscle degeneration in ALS. *TNFRSF12*, also known as *TWEAK*, is a cytokine that promotes apoptosis, chronic inflammation, and cell-specific angiogenesis. It is involved in adult neurogenesis, synaptic functions, and skeletal muscle atrophy and regeneration[121,122], and is expressed in the central nervous system. Although *TWEAK* was elevated in ALS patients compared to controls, no clear temporal trend was observed after diagnosis and no significant difference was found in CSF. Myoglobin, *MB*, has not been studied before as a diagnostic biomarker for ALS, and further studies are required to confirm the specificity of *MB* in diagnosing ALS. The association between higher CSF *MB* levels and a higher risk of death after ALS diagnosis suggests *MB* is a new prognostic indicator for ALS. However, more studies are needed to confirm this finding and to understand the relationship between *MB* levels in circulation and CSF. *GDF-15* belongs to the transforming growth factor *TGF-β* superfamily and is a macrophage inhibitory cytokine[123,124]. Our study revealed the potential of *GDF-15* as a CSF biomarker for ALS diagnosis and a plasma biomarker for predicting functional decline following ALS diagnosis. The increasing plasma level of *GDF-15* over time since ALS diagnosis suggests its involvement in chronic

inflammatory pathways in ALS progression[125]. The findings of *GDF-15* may also be linked to altered metabolism and weight loss in ALS, as *GDF-15* has previously been implicated in these processes[126-129].

Besides the *potential risk proteins*, the study also revealed *potential protective proteins* including *RGMA, GDF-8, MMP-3, IL32, CDH6, ITGB2,* and *GCP5* that a lower level of protein plasma was associated with a higher risk of death after ALS diagnosis and ALS patients demonstrated a decreasing level of these proteins in plasma by time since diagnosis. *RGMA* or repulsive guidance molecule bone morphogenetic protein co-receptor A has been shown to inhibit neuronal regeneration and neuroprotection in multiple sclerosis[130-133] and has been implicated in Parkinson's disease as *RGMA* over-expression might induce progressive movement disorders[134]. In contrast to *RGMA*, which is seldom investigated in ALS, several studies have suggested the role of *GDF-8* (also known as myostatin) in muscle atrophy in ALS and other diseases[135-143]. *IL32* or interleukin-32 is a proinflammatory cytokine [144] that induces cytokine production such as *TNF-α* (involved in *p38 MAPK* pathway) and *IL8* from macrophages[121,145]. Studies have also shown that *IL32 expression was down-regulated in spinal muscular atrophy (SMA) patients[146] and IL32* could activate *NF-kB* signaling via the *p38 MAPK* pathway[147,148]. *CDH6*, cadherin-6, is a cell-cell adhesion molecule[149,150] to maintain extracellular domain integrity and was previously shown to be enriched in spinal motor neurons [151]. Our finding of a lower risk of death in ALS patients with a higher level of *CDH6* suggests that higher expression of cadherin may promote cell-cell adhesion in the extracellular domain, leading to the integrity and maintenance of the blood-brain barrier (BBB)[152]. Adversely, decreasing levels of *CDH6* over time after ALS diagnosis might indicate a continued loss of BBB integrity. *ITGB2* or integrin beta 2 is essential for cell adhesion, leukocyte trafficking, T-cell activation, phagocytosis, etc.[153]. *GCP5*, tubulin gamma complex associated protein 5, is on the other hand required for microtubule nucleation[154] and is highly expressed in skeletal muscles and brain[155]. The reorganization of microtubule nucleation is essential for muscle differentiation[156]. Therefore, the link between *GCP5* and functional decline could potentially indicate a lower degree of muscle differentiation and thus a lower functional ability in ALS patients when compared to controls and over time since diagnosis.

Other proteins including *FGF-21, CHIT1,* and *GPNMB* were also indicative of ALS risk. ALS patients exhibited a higher level of *FGF-21* in plasma compared to controls. Mitochondrial dysfunction is an early pathophysiological event in ALS[157], and in response to this, neurons may release *FGF-21*[158]. While it is unclear why the elevated level was only observed in plasma and not CSF, research has shown that there is a higher level of *FGF-21* in plasma than in CSF in healthy individuals[159]. Additionally, ALS patients showed higher levels of *CHIT1* and *GPNMB* in CSF compared to the controls. *CHIT1* in CSF has previously been shown to have high specificity in separating ALS patients from healthy controls and those with other neurodegenerative diseases[160-163] and is involved in macrophage activation[160] which in turn influences the function of microglial and decelerates ALS disease progression[164].

For Study III, the study population is representative of the Stockholm ALS population during the conduct of the ALSrisc study and has considered the assessment of both plasma and CSF samples from the ALS patients as well as the controls. The use of proteomics profiling by Olink Proteomics technology allows for high sensitivity and detection of low protein concentrations. Additionally, this study is the first to examine longitudinal trajectories of protein biomarkers in both plasma and CSF in ALS patients. On the other hand, the small number of patients with *C9orf72* mutations made it challenging to stratify and analyze this group separately, and it is unclear if there are specific protein biomarkers for *C9orf72*-related ALS. We used sibling and spouse controls to compare with ALS patients but did not have access to population-based controls, which may have led to an underestimation of the differences between ALS patients and disease-free individuals. However, similar results observed when using sibling and spouse controls provided some reassurance. As ALS is a rapidly progressing disease and our study included mostly incident patients, caution should be taken when interpreting the results from the longitudinal trajectories analyses of the proteins. Finally, it is uncertain if our findings are generalizable to ALS patients outside of Sweden.

7 POINTS OF PERSPECTIVE

The vibrant development of bulk and single-cell omics technologies left limitations that create opportunities for improvements. The thesis forged to understand some of the limitations in the technologies, and discovered novel genetic and proteomics signatures specific to brain regions and ALS respectively.

Study I focused on the development of an isoform quantification method for droplet-based scRNA-Seq technologies to enhance accuracies in isoform quantification by addressing the prime-bias problem introduced by these technologies. The method could be further assessed and applied to a wider range of sequencing methods. Isoform-specific downstream analyses could also be incorporated into the software to expand beyond the current quantification step.

Study II aimed to examine and discover novel *cis*-splicing variants based on the isoform-ratio patterns within each gene. Further studies could be done to examine the downstream effects of the irQTL in other omics types.

Study III identified novel protein biomarkers specific to ALS diagnosis, prognosis, functional decline, and longitudinal proteins associated with disease progression. Further studies are needed to validate such findings, especially if these biomarkers are specific to the ALS disease, or rather they are also commonly found in other neurodegenerative diseases.

8 ACKNOWLEDGEMENTS

I would like to thank all participants, project leaders, and coordinators of the ALSrisc study. Without your contributions, time, and effort, the research for Study III would not have been possible. I would also like to acknowledge that the computations and data handling of the thesis projects were enabled by the resources at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Trung Nghia Vu: my main supervisor. I would like to thank you for all your support and guidance during my time here. Your guidance and feedback have contributed significantly to the success of the thesis work. Thank you for always being available to answer my questions and pushing the projects forward. I feel lucky to have the opportunity to be your student, and I will always be grateful for you being my supervisor during my time here.

Fang Fang: my co-supervisor. I felt fortunate to have you as my co-supervisor. You are like a guiding angel, offering all kinds of support, guidance, and encouragement whenever you can. Your dedication has been a source of strength for me and has inspired me to be a person like you. Thank you for believing in me and I will always remember you.

Yudi Pawitan: my co-supervisor. I wanted to take a moment to express my appreciation for the time and effort you have invested in the projects. Your expertise has been valuable and I appreciate the effort and feedback you have put into my work. I know that you hold high expectations for your team, and I am grateful for the ways you have pushed me to improve and grow.

Xia Shen: my co-supervisor. I wanted to thank you for your time and effort in Study II of my thesis work. I wanted to acknowledge the opportunity you have offered to me and the lessons learned from those difficult situations.

Paolo Parini: my mentor. I wanted to express my deep appreciation for all the advice and guidance you have given me so far. The altitude of how you perceive career and life is what I wanted to learn from you. I am so grateful for the ways you have helped me in improving myself. I felt fortunate to have made the right choice to have you as my mentor. You are truly a wonderful person. Thank you so much for being an inspiring role model and friend. I will always treasure the advice and experience you have shared with me.

Vladimir Kuznetsov: my bachelor supervisor. I always felt very fortunate for having you as my supervisor, and I always remember all the knowledge you have taught me so far. You have truly made a significant impact on my research development and I was always amazed and thrilled by your thoughts and ideas, especially when you transformed physics concepts into mathematical models for biological data. Your guidance and support brought gigantic steps in helping me to achieve my goals in research. Your width and depth of knowledge and your willingness to share your years of research experience with me have shaped me to think critically and think in a multi-perspective way while doing research. You are the first person

when I stood by the bus stop at seven in the morning and decided that, “I wanted to be a scientist, and do research just like you”. Without you, I would not have come so far.

Andrey Alexeyenko: my previous supervisor at SciLifeLab. I wanted to express my gratitude for your guidance in my research journey. Even though it was a short time and I have to leave quickly for my graduation, your guidance in research has made me understand the purpose of what I was learning back then. The opportunity of working with you has guided my research direction.

Marie Jansson: you could bring happiness to people around you. Thank you, Marie, for all your support and wonderful tips at work throughout the past years. Thank you for everything you have done for us, and for being always so supportive, and I am always been encouraged by your smile and your positivity at work whenever I see you. You are truly an asset to the Biostatistics group and I felt very fortunate to have you around.

Mark Clements: thank you so much for helping me in solving so many statistical questions throughout the years at MEB. I wanted to thank you from my heart and I appreciate the time and effort you have put in to help me whenever you can. You are a great professor and I felt very lucky to have you around.

Alessandra Nanni: I wanted to thank you for all your assistance over the past few years at MEB. You have done the best you can in helping me in addressing all my concerns.

Christina Seitz: thank you for all the help in Study III. You have always been helpful and supportive over the past few years, and I am happy to have worked with you.

I would like to thank my collaborators, Anders Malarstig, Åsa Hedman, John Andersson, Caroline Ingre, Yan Chen, Aniko Lovik, Caroline Graff, and Jose Laffita for your guidance and support in making the success out of my thesis work. I would also like to thank Charilaos Chourpiliadis, Chenqing Zheng, Quang Thinh Trac, Emily Joyce, Abbe Ullgren, Aniko Lovik, Kristin Samuelsson, Rayomand Press, and Fredrik Piehl for their contributions to my thesis projects.

I would like to thank my colleagues and the professors at the department of Medical Epidemiology and Biostatistics, especially Zheng Ning, Gabriel Isheden, Weiyao Yin, Nikolaos Skourlis, Yuying Li, Quang Thinh Trac, Ziyan Ma, Tong Gong, Enoch Yi-Tung Chen, Marie Reilly, Gustav Jonzon, Alessandra Nanni, Gunilla Sonnerbring, Wenjiang Deng, Therese Andersson, Arvid Sjölander, Keith Humphreys, Alex Ploner, Yuliya Leontyeva, and many others from the Biostatistics and Epidemiology groups. I would also like to thank the IT support for their assistance over the past years.

To all my friends at the Institutet för Miljömedicin: Kejia Hu, Hang Yu, Jianing Liu, Yihui Yang, Yihan Hu, Shifeng Lian, Emily Joyce, Christina Seitz, Charilaos Chourpiliadis, John Andersson, Jing Zhou, Dang Wei, Donghao Lu, Jacob Bergstedt, Fatemeh Sadeghi, Marion Opatowski, Mary Barker, Elgeta Hysaj, Elie Diba, Emma Bränn, Jiangwei Sun, Qian Yang,

and all others at the corridor, thank you for all your support and wonderful times we spent together!

I wanted to thank Xuexin Li, Paolo Parini, Roman Tremmel, Joseph Loscalzo, Volker Lauschke, Bradley Maron, Paola Paci, Ingemar Ernberg, Nguan Soon Tan, Zehuan Liao, Weiyao Yin for all your support, guidance, and advice in research.

至我亲爱的母亲，爷爷，奶奶，姨姨和姨丈: 非常感谢你们长期以来对我无条件的支持和爱。如果没有你们，我也不走不到今天这一步。是你们让我，在出了社会以后，在我对面生活困难的时候，学会积极向上，学会跌倒了自己爬起来，咬紧牙关继续往前走。你们，是我前进的动力。我爱你们！

I felt fortunate to have all of you involved in my life and shaping what I am today. Life is a long journey, and we are at a cross-road now. Take care, everyone.

9 REFERENCES

1 Lu Pan, C. Z., Zhijian Yang, Yudi Pawitan, Trung Nghia Vu, Xia Shen. Hidden genetic regulation of human complex traits via brain isoforms. *Phenomics* (In press).

2 Lu Pan, C. S., Caroline Ingre, Åsa Hedman, Jose Laffita, Trung Nghia Vu, Yudi Pawitan, Abbe Ullgren, Solmaz Yazdani, John Andersson, Emily Joyce, Aniko Lovik, Yan Chen, Kristin Samuelsson, Rayomand Press, Fredrik Piehl, Caroline Graff, Anders Mälarstig, Fang Fang. Protein biomarkers in risk and prognosis of amyotrophic lateral sclerosis. (Manuscript in preparation).

3 Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561-563 (1970). https://doi.org:10.1038/227561a0

4 Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18**, 83 (2017). https://doi.org:10.1186/s13059-017-1215-1

5 Yadav, S. P. The wholeness in suffix-omics,-omes, and the word om. *Journal of biomolecular techniques: JBT* **18**, 277 (2007).

6 Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nature Methods* **17**, 11-14 (2020). https://doi.org:10.1038/s41592-019-0691-5

7 Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature Reviews Genetics* **19**, 299-310 (2018). https://doi.org:10.1038/nrg.2018.4

8 Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56-68 (2011). https://doi.org:10.1038/nrg2918

9 Gibson, G. Decanalization and the origin of complex disease. *Nature Reviews Genetics* **10**, 134-140 (2009). https://doi.org:10.1038/nrg2502

10 Kreitmaier, P., Katsoula, G. & Zeggini, E. Insights from multi-omics integration in complex disease primary tissues. *Trends in Genetics* **39**, 46-58 (2023). https://doi.org:10.1016/j.tig.2022.08.005

11 Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology* **38**, 737-746 (2020). https://doi.org:10.1038/s41587-020-0465-8

12 He, X., Memczak, S., Qu, J., Belmonte, J. C. I. & Liu, G.-H. Single-cell omics in ageing: a young and growing field. *Nature Metabolism* **2**, 293-302 (2020). https://doi.org:10.1038/s42255-020-0196-7

13 Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across modalities. *Nature Methods* **17**, 14-17 (2020). https://doi.org:10.1038/s41592-019-0692-4

14 Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative Methods and Practical Challenges for Single-Cell Multi-omics. *Trends in Biotechnology* **38**, 1007-1022 (2020). https://doi.org:https://doi.org/10.1016/j.tibtech.2020.02.013

15 Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics, Proteomics & Bioinformatics* **19**, 253-266 (2021). https://doi.org:https://doi.org/10.1016/j.gpb.2020.02.005

16 Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*, 2021.2005.2005.442755 (2021). https://doi.org:10.1101/2021.05.05.442755

17 Hu, Y., Wang, K. & Li, M. Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers. *PLoS Comput Biol* **16**, e1007925 (2020). https://doi.org:10.1371/journal.pcbi.1007925

18 Huang, Y. & Sanguinetti, G. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology* **18**, 123 (2017). https://doi.org:10.1186/s13059-017-1248-5

19 Song, Y. *et al.* Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Mol Cell* **67**, 148-161.e145 (2017). https://doi.org:10.1016/j.molcel.2017.06.003

20 Rowland, L. P. & Shneider, N. A. Amyotrophic Lateral Sclerosis. *New England Journal of Medicine* **344**, 1688-1700 (2001). https://doi.org:10.1056/nejm200105313442207

21 Feldman, E. L. *et al.* Amyotrophic lateral sclerosis. *The Lancet* **400**, 1363-1380 (2022). https://doi.org:https://doi.org/10.1016/S0140-6736(22)01272-7

22 Buescher, J. M. & Driggers, E. M. Integration of omics: more than the sum of its parts. *Cancer & Metabolism* **4**, 4 (2016). https://doi.org:10.1186/s40170-016-0143-y

23 Dahm, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology* **278**, 274-288 (2005). https://doi.org:https://doi.org/10.1016/j.ydbio.2004.11.028

24 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953). https://doi.org:10.1038/171737a0

25 Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441-448 (1975). https://doi.org:https://doi.org/10.1016/0022-2836(75)90213-2

26 Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560-564 (1977). https://doi.org:10.1073/pnas.74.2.560

27 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977). https://doi.org:10.1073/pnas.74.12.5463

28 Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1-8 (2016). https://doi.org:10.1016/j.ygeno.2015.11.003

29 Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345-353 (2017). https://doi.org:10.1038/nature24286

30 Potapov, V. & Ong, J. L. Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One* **12**, e0169774 (2017). https://doi.org:10.1371/journal.pone.0169774

31 Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Human Molecular Genetics* **19**, R227-R240 (2010). https://doi.org:10.1093/hmg/ddq416

32 Xiao, T. & Zhou, W. The third generation sequencing: the advanced approach to genetic diseases. *Transl Pediatr* **9**, 163-173 (2020). https://doi.org:10.21037/tp.2020.03.06

33 Bleidorn, C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity* **14**, 1-8 (2016). https://doi.org:10.1080/14772000.2015.1099575

34 Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nature Medicine* **28**, 243-250 (2022). https://doi.org:10.1038/s41591-021-01672-4

35 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001). https://doi.org:10.1038/35057062

36 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022). https://doi.org:doi:10.1126/science.abj6987

37 Collins, F. S., Doudna, J. A., Lander, E. S. & Rotimi, C. N. Human Molecular Genetics and Genomics — Important Advances and Exciting Possibilities. *New England Journal of Medicine* **384**, 1-4 (2021). https://doi.org:10.1056/NEJMp2030694

38 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015). https://doi.org:10.1371/journal.pmed.1001779

39 Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 59 (2021). https://doi.org:10.1038/s43586-021-00056-9

40 Akiyama, M. Multi-omics study for interpretation of genome-wide association study. *Journal of Human Genetics* **66**, 3-10 (2021). https://doi.org:10.1038/s10038-020-00842-5

41 Milward, E. A. *et al.* in *Encyclopedia of Cell Biology* (eds Ralph A. Bradshaw & Philip D. Stahl) 160-165 (Academic Press, 2016).

42 Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine* **52**, 1452-1465 (2020). https://doi.org:10.1038/s12276-020-0422-0

43 Tahiliani, J. *et al.* Utility of RNA Sequencing Analysis in the Context of Genetic Testing. *Current Genetic Medicine Reports* **8**, 140-146 (2020). https://doi.org:10.1007/s40142-020-00195-7

44 Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009). https://doi.org:10.1038/nmeth.1315

45 Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nature Communications* **11**, 4307 (2020). https://doi.org:10.1038/s41467-020-18158-5

46 Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harbor Protocols* **2015**, pdb.top084970 (2015). https://doi.org:10.1101/pdb.top084970

47 Fields, S. Proteomics in Genomeland. *Science* **291**, 1221-1224 (2001). https://doi.org:doi:10.1126/science.291.5507.1221

48 Patterson, S. D. & Aebersold, R. H. Proteomics: the first decade and beyond. *Nature Genetics* **33**, 311-323 (2003). https://doi.org:10.1038/ng1106

49 Belczacka, I. *et al.* Proteomics biomarkers for solid tumors: Current status and future prospects. *Mass spectrometry reviews* **38**, 49-78 (2019).

50 Mischak, H., Delles, C., Vlahou, A. & Vanholder, R. Proteomic biomarkers in kidney disease: issues in development and implementation. *Nature Reviews Nephrology* **11**, 221-232 (2015).

51 Frantzi, M. *et al.* Developing proteomic biomarkers for bladder cancer: towards clinical application. *Nature Reviews Urology* **12**, 317-330 (2015).

52 Frantzi, M., Bhat, A. & Latosinska, A. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clinical and translational medicine* **3**, 1-22 (2014).

53 Anderson, L. Candidate‐based proteomics in the search for biomarkers of cardiovascular disease. *The Journal of physiology* **563**, 23-60 (2005).

54 Kisluk, J., Ciborowski, M., Niemira, M., Kretowski, A. & Niklinski, J. Proteomics biomarkers for non-small cell lung cancer. *Journal of Pharmaceutical and Biomedical Analysis* **101**, 40-49 (2014).

55 Khoo, A. *et al.* Proteomic discovery of non-invasive biomarkers of localized prostate cancer using mass spectrometry. *Nature Reviews Urology* **18**, 707-724 (2021). https://doi.org:10.1038/s41585-021-00500-1

56 Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature Biotechnology* **24**, 971-983 (2006). https://doi.org:10.1038/nbt1235

57 Park, S. A., Han, S. M. & Kim, C. E. New fluid biomarkers tracking non-amyloid-β and non-tau pathology in Alzheimer's disease. *Experimental & Molecular Medicine* **52**, 556-568 (2020). https://doi.org:10.1038/s12276-020-0418-9

58 Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16**, 133-145 (2015). https://doi.org:10.1038/nrg3833

59 Chandra, V. *et al.* Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. *Nature Genetics* **53**, 110-119 (2021). https://doi.org:10.1038/s41588-020-00745-3

60 Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580-585 (2013). https://doi.org:10.1038/ng.2653

61 Yao, C. *et al.* Genome‐wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications* **9**, 3268 (2018). https://doi.org:10.1038/s41467-018-05512-x

62 Walker, R. L. *et al.* Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. *Cell* **179**, 750-771.e722 (2019). https://doi.org:10.1016/j.cell.2019.09.021

63 Zhang, Y. *et al.* Regional Variation of Splicing QTLs in Human Brain. *Am J Hum Genet* **107**, 196-210 (2020). https://doi.org:10.1016/j.ajhg.2020.06.002

64 Wang, Q. *et al.* Longitudinal data in peripheral blood confirm that PM20D1 is a quantitative trait locus (QTL) for Alzheimer's disease and implicate its dynamic role in disease progression. *Clinical Epigenetics* **12**, 189 (2020). https://doi.org:10.1186/s13148-020-00984-5

65 Humphrey, J. *et al.* Integrative genetic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. *medRxiv*, 2021.2008.2031.21262682 (2021). https://doi.org:10.1101/2021.08.31.21262682

66 van Es, M. A. *et al.* Amyotrophic lateral sclerosis. *Lancet* **390**, 2084-2098 (2017). https://doi.org:10.1016/s0140-6736(17)31287-4

67 Cady, J. *et al.* Amyotrophic lateral sclerosis onset is influenced by the burden of rare variants in known amyotrophic lateral sclerosis genes. *Ann Neurol* **77**, 100-113 (2015). https://doi.org:10.1002/ana.24306

68 Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat Neurosci* **17**, 17-23 (2014). https://doi.org:10.1038/nn.3584

69 Estevez-Fraga, C. *et al.* Expanding the Spectrum of Movement Disorders Associated With C9orf72 Hexanucleotide Expansions. *Neurology Genetics* **7**, e575 (2021). https://doi.org:10.1212/nxg.0000000000000575

70 Cooper-Knock, J. *et al.* C9ORF72 expansions, parkinsonism, and Parkinson disease: a clinicopathologic study. *Neurology* **81**, 808-811 (2013). https://doi.org:10.1212/WNL.0b013e3182a2cc38

71 Hensman Moss, D. J. *et al.* C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. *Neurology* **82**, 292-299 (2014). https://doi.org:10.1212/wnl.0000000000000061

72 McCann, E. P. *et al.* Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. *J Med Genet* (2020). https://doi.org:10.1136/jmedgenet-2020-106866

73 Al-Chalabi, A. & Hardiman, O. The epidemiology of ALS: a conspiracy of genes, environment and time. *Nat Rev Neurol* **9**, 617-628 (2013). https://doi.org:10.1038/nrneurol.2013.203

74 Mitchell, J. D. & Borasio, G. D. Amyotrophic lateral sclerosis. *Lancet* **369**, 2031-2041 (2007). https://doi.org:10.1016/s0140-6736(07)60944-1

75 Bruening, G. & Lyons, J. The case of the FLAVR SAVR tomato. *California Agriculture* **54**, 6-7 (2000).

76 Hardiman, O. *et al.* Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers* **3**, 17071 (2017). https://doi.org:10.1038/nrdp.2017.71

77 Turner, M. R., Kiernan, M. C., Leigh, P. N. & Talbot, K. Biomarkers in amyotrophic lateral sclerosis. *Lancet Neurol* **8**, 94-109 (2009). https://doi.org:10.1016/s1474-4422(08)70293-x

78 Robberecht, W. & Philips, T. The changing scene of amyotrophic lateral sclerosis. *Nature Reviews Neuroscience* **14**, 248-264 (2013). https://doi.org:10.1038/nrn3430

79 Frazee, A. C., Jaffe, A. E., Langmead, B. & Leek, J. T. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778-2784 (2015). https://doi.org:10.1093/bioinformatics/btv272

80 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e1821 (2019). https://doi.org:10.1016/j.cell.2019.05.031

81 Ranzoni, A. M. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28**, 472-487.e477 (2021). https://doi.org:https://doi.org/10.1016/j.stem.2020.11.015

82 Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648-660 (2015). https://doi.org:doi:10.1126/science.1262110

83 Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851-4853 (2019). https://doi.org:10.1093/bioinformatics/btz469

84 DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245-256 (2011). https://doi.org:10.1016/j.neuron.2011.09.011

85 Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology* **32**, 462-464 (2014). https://doi.org:10.1038/nbt.2862

86 Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L. & Tse, D. N. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biology* **17**, 112 (2016). https://doi.org:10.1186/s13059-016-0970-8

87 Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm? *Nature Biotechnology* **26**, 897-899 (2008). https://doi.org:10.1038/nbt1406

88 Deng, W. *et al.* Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data. *Bioinformatics* **36**, 805-812 (2019). https://doi.org:10.1093/bioinformatics/btz640

89 Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biology* **20**, 65 (2019). https://doi.org:10.1186/s13059-019-1670-y

90 Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39**, 813-818 (2021). https://doi.org:10.1038/s41587-021-00870-2

91 Pan, L., Dinh, H. Q., Pawitan, Y. & Vu, T. N. Isoform-level quantification for single-cell RNA sequencing. *Bioinformatics* **38**, 1287-1294 (2021). https://doi.org:10.1093/bioinformatics/btab807

92 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007). https://doi.org:10.1086/519795

93 Haller, T., Kals, M., Esko, T., Mägi, R. & Fischer, K. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief Bioinform* **16**, 39-44 (2015). https://doi.org:10.1093/bib/bbt066

94 Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017). https://doi.org:10.1093/bioinformatics/btw613

95 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291-295 (2015). https://doi.org:10.1038/ng.3211

96 Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature Genetics* **49**, 1421-1427 (2017). https://doi.org:10.1038/ng.3954

97 Longinetti, E. *et al.* The Swedish motor neuron disease quality registry. *Amyotroph Lateral Scler Frontotemporal Degener* **19**, 528-537 (2018). https://doi.org:10.1080/21678421.2018.1497065

98 LIANG, K.-Y. & ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22 (1986). https://doi.org:10.1093/biomet/73.1.13

99 Cedarbaum, J. M. *et al.* The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). *J Neurol Sci* **169**, 13-21 (1999). https://doi.org:10.1016/s0022-510x(99)00210-5

100 Lausen, B. & Schumacher, M. Maximally Selected Rank Statistics. *Biometrics* **48**, 73-85 (1992). https://doi.org:10.2307/2532740

101 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525-527 (2016). https://doi.org:10.1038/nbt.3519

102 Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017). https://doi.org:10.1038/ncomms14049

103 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417-419 (2017). https://doi.org:10.1038/nmeth.4197

104 Sarkar, H., Srivastava, A., Bravo, H. C., Love, M. I. & Patro, R. Terminus enables the discovery of data-driven, robust transcript groups from RNA-seq data. *Bioinformatics* **36**, i102-i110 (2020). https://doi.org:10.1093/bioinformatics/btaa448

105 Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications* **5**, 4698 (2014). https://doi.org:10.1038/ncomms5698

106 Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nature Genetics* **54**, 1355-1363 (2022). https://doi.org:10.1038/s41588-022-01154-4

107 Mayer, G., Kröger, M. & Meier-Ewert, K. Effects of vitamin B12 on performance and circadian rhythm in normal subjects. *Neuropsychopharmacology* **15**, 456-464 (1996). https://doi.org:10.1016/s0893-133x(96)00055-3

108 Hashimoto, S. *et al.* Vitamin B12 enhances the phase-response of circadian melatonin rhythm to a single bright light exposure in humans. *Neuroscience Letters* **220**, 129-132 (1996). https://doi.org:https://doi.org/10.1016/S0304-3940(96)13247-X

109 Safran, M. *et al.* in *Practical Guide to Life Science Databases* (eds Imad Abugessaisa & Takeya Kasukawa) 27-56 (Springer Nature Singapore, 2021).

110 Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Research* **50**, D988-D995 (2021). https://doi.org:10.1093/nar/gkab1049

111 Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Research* **49**, D916-D923 (2020). https://doi.org:10.1093/nar/gkaa1087

112 Lu, C.-H. *et al.* Neurofilament light chain. *A prognostic biomarker in amyotrophic lateral sclerosis* **84**, 2247-2257 (2015). https://doi.org:10.1212/wnl.0000000000001642

113 Gaetani, L. *et al.* Neurofilament light chain as a biomarker in neurological disorders. *Journal of Neurology, Neurosurgery & Psychiatry* **90**, 870-881 (2019). https://doi.org:10.1136/jnnp-2018-320106

114 Verde, F., Otto, M. & Silani, V. Neurofilament Light Chain as Biomarker for Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Frontiers in Neuroscience* **15** (2021). https://doi.org:10.3389/fnins.2021.679199

115 Khalil, M. *et al.* Neurofilaments as biomarkers in neurological disorders. *Nature Reviews Neurology* **14**, 577-589 (2018). https://doi.org:10.1038/s41582-018-0058-z

116 Byrne, L. M. *et al.* Neurofilament light protein in blood as a potential biomarker of neurodegeneration in Huntington's disease: a retrospective cohort analysis. *The Lancet Neurology* **16**, 601-609 (2017). https://doi.org:https://doi.org/10.1016/S1474-4422(17)30124-2

117 Dreger, M. *et al.* Cerebrospinal Fluid Neurofilament Light Chain (NfL) Predicts Disease Aggressiveness in Amyotrophic Lateral Sclerosis: An Application of the D50 Disease Progression Model. *Frontiers in Neuroscience* **15** (2021). https://doi.org:10.3389/fnins.2021.651651

118 Verde, F. *et al.* Neurofilament light chain in serum for the diagnosis of amyotrophic lateral sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* **90**, 157-164 (2019). https://doi.org:10.1136/jnnp-2018-318704

119 Harris, S. E. *et al.* Neurology-related protein biomarkers are associated with cognitive ability and brain volume in older age. *Nature Communications* **11**, 800 (2020). https://doi.org:10.1038/s41467-019-14161-7

120 Menni, C. *et al.* Circulating proteomic signatures of chronological age. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* **70**, 809-816 (2015).

121 Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). https://doi.org:doi:10.1126/science.1260419

122 Burkly, L. C. TWEAK/Fn14 axis: The current paradigm of tissue injury-inducible function in the midst of complexities. *Seminars in Immunology* **26**, 229-236 (2014). https://doi.org:https://doi.org/10.1016/j.smim.2014.02.006

123 Bootcov, M. R. *et al.* MIC-1, a novel macrophage inhibitory cytokine, is a divergent member of the TGF-beta superfamily. *Proc Natl Acad Sci U S A* **94**, 11514-11519 (1997). https://doi.org:10.1073/pnas.94.21.11514

124 Wischhusen, J., Melero, I. & Fridman, W. H. Growth/Differentiation Factor-15 (GDF-15): From Biomarker to Novel Targetable Immune Checkpoint. *Front Immunol* **11**, 951 (2020). https://doi.org:10.3389/fimmu.2020.00951

125 Luan, H. H. *et al.* GDF15 Is an Inflammation-Induced Central Mediator of Tissue Tolerance. *Cell* **178**, 1231-1244.e1211 (2019). https://doi.org:10.1016/j.cell.2019.07.033

126 Emmerson, P. J. *et al.* The metabolic effects of GDF15 are mediated by the orphan receptor GFRAL. *Nat Med* **23**, 1215-1219 (2017). https://doi.org:10.1038/nm.4393

127 Hsu, J. Y. *et al.* Non-homeostatic body weight regulation through a brainstem-restricted receptor for GDF15. *Nature* **550**, 255-259 (2017). https://doi.org:10.1038/nature24042

128 Yang, L. *et al.* GFRAL is the receptor for GDF15 and is required for the anti-obesity effects of the ligand. *Nat Med* **23**, 1158-1166 (2017). https://doi.org:10.1038/nm.4394

129 L'Homme, L. *et al.* Saturated Fatty Acids Promote GDF15 Expression in Human Macrophages through the PERK/eIF2/CHOP Signaling Pathway. *Nutrients* **12** (2020). https://doi.org:10.3390/nu12123771

130 Demicheva, E. *et al.* Targeting repulsive guidance molecule A to promote regeneration and neuroprotection in multiple sclerosis. *Cell Rep* **10**, 1887-1898 (2015). https://doi.org:10.1016/j.celrep.2015.02.048

131 Tanabe, S. & Yamashita, T. Repulsive guidance molecule-a is involved in Th17-cell-induced neurodegeneration in autoimmune encephalomyelitis. *Cell Rep* **9**, 1459-1470 (2014). https://doi.org:10.1016/j.celrep.2014.10.038

132 Kubo, T., Tokita, S. & Yamashita, T. Repulsive guidance molecule-a and demyelination: implications for multiple sclerosis. *J Neuroimmune Pharmacol* **7**, 524-528 (2012). https://doi.org:10.1007/s11481-011-9334-z

133 Muramatsu, R. *et al.* RGMa modulates T cell responses and is involved in autoimmune encephalomyelitis. *Nat Med* **17**, 488-494 (2011). https://doi.org:10.1038/nm.2321

134 Korecka, J. A. *et al.* Repulsive Guidance Molecule a (RGMa) Induces Neuropathological and Behavioral Changes That Closely Resemble Parkinson's Disease. *The Journal of Neuroscience* **37**, 9361-9379 (2017). https://doi.org:10.1523/jneurosci.0084-17.2017

135 Holzbaur, E. L. F. *et al.* Myostatin inhibition slows muscle atrophy in rodent models of amyotrophic lateral sclerosis. *Neurobiology of Disease* **23**, 697-707 (2006). https://doi.org:https://doi.org/10.1016/j.nbd.2006.05.009

136 Walsh, F. S. & Celeste, A. J. Myostatin: a modulator of skeletal-muscle stem cells. *Biochemical Society Transactions* **33**, 1513-1517 (2005). https://doi.org:10.1042/bst0331513

137 Sakuma, K. & Yamaguchi, A. Inhibitors of Myostatin- and Proteasome-Dependent Signaling for Attenuating Muscle Wasting. *Recent Patents on Regenerative Medicine* **1**, 284-298 (2011).

138 Lee, J. H. & Jun, H.-S. Role of Myokines in Regulating Skeletal Muscle Mass and Function. *Frontiers in Physiology* **10** (2019). https://doi.org:10.3389/fphys.2019.00042

139 Sumner, C. J. *et al.* Inhibition of myostatin does not ameliorate disease features of severe spinal muscular atrophy mice. *Human Molecular Genetics* **18**, 3145-3152 (2009). https://doi.org:10.1093/hmg/ddp253

140 Smith, R. C. & Lin, B. K. Myostatin inhibitors as therapies for muscle wasting associated with cancer and other disorders. *Curr Opin Support Palliat Care* **7**, 352-360 (2013). https://doi.org:10.1097/spc.0000000000000013

141 Walker, R. G. *et al.* Molecular characterization of latent GDF8 reveals mechanisms of activation. *Proceedings of the National Academy of Sciences* **115**, E866-E875 (2018). https://doi.org:doi:10.1073/pnas.1714622115

142 Mariot, V. *et al.* Downregulation of myostatin pathway in neuromuscular diseases may explain challenges of anti-myostatin therapeutic approaches. *Nature Communications* **8**, 1859 (2017). https://doi.org:10.1038/s41467-017-01486-4

143 Liu, C. M. *et al.* Myostatin antisense RNA-mediated muscle growth in normal and cancer cachexia mice. *Gene Therapy* **15**, 155-160 (2008). https://doi.org:10.1038/sj.gt.3303016

144 Dinarello, C. A. & Kim, S.-H. IL-32, a novel cytokine with a possible role in disease. *Annals of the Rheumatic Diseases* **65**, iii61-iii64 (2006). https://doi.org:10.1136/ard.2006.058511

145 Kim, S. Interleukin-32 in inflammatory autoimmune diseases. *Immune Netw* **14**, 123-127 (2014). https://doi.org:10.4110/in.2014.14.3.123

146 Papadimitriou, D. *et al.* Inflammation in ALS and SMA: sorting out the good from the evil. *Neurobiol Dis* **37**, 493-502 (2010). https://doi.org:10.1016/j.nbd.2009.10.005

147 Aass, K. R., Kastnes, M. H. & Standal, T. Molecular interactions and functions of IL-32. *Journal of Leukocyte Biology* **109**, 143-159 (2021). https://doi.org:https://doi.org/10.1002/JLB.3MR0620-550R

148 Xin, T., Chen, M., Duan, L., Xu, Y. & Gao, P. Interleukin-32: its role in asthma and potential as a therapeutic agent. *Respiratory Research* **19**, 124 (2018). https://doi.org:10.1186/s12931-018-0832-x

149 Shapiro, L. *et al.* Structural basis of cell-cell adhesion by cadherins. *Nature* **374**, 327-337 (1995). https://doi.org:10.1038/374327a0

150 Perez, T. D. & Nelson, W. J. Cadherin adhesion: mechanisms and molecular interactions. *Handb Exp Pharmacol*, 3-21 (2004). https://doi.org:10.1007/978-3-540-68170-0_1

151 Allodi, I. *et al.* Modeling Motor Neuron Resilience in ALS Using Stem Cells. *Stem Cell Reports* **12**, 1329-1341 (2019). https://doi.org:https://doi.org/10.1016/j.stemcr.2019.04.009

152 Li, W., Chen, Z., Chin, I., Chen, Z. & Dai, H. The Role of VE-cadherin in Blood-brain Barrier Integrity Under Central Nervous System Pathological Conditions. *Curr Neuropharmacol* **16**, 1375-1384 (2018). https://doi.org:10.2174/1570159x16666180222164809

153 Fagerholm, S. C., Guenther, C., Llort Asens, M., Savinko, T. & Uotila, L. M. Beta2-Integrins and Interacting Proteins in Leukocyte Trafficking, Immune Suppression, and Immunodeficiency Disease. *Front Immunol* **10**, 254 (2019). https://doi.org:10.3389/fimmu.2019.00254

154 Murphy, S. M. *et al.* GCP5 and GCP6: two new members of the human gamma-tubulin complex. *Mol Biol Cell* **12**, 3340-3352 (2001). https://doi.org:10.1091/mbc.12.11.3340

155 Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics* **54**, 1.30.31-31.30.33 (2016). https://doi.org:https://doi.org/10.1002/cpbi.5

156 Bugnard, E., Zaal, K. J. & Ralston, E. Reorganization of microtubule nucleation during muscle differentiation. *Cell Motil Cytoskeleton* **60**, 1-13 (2005). https://doi.org:10.1002/cm.20042

157 Smith, E. F., Shaw, P. J. & De Vos, K. J. The role of mitochondria in amyotrophic lateral sclerosis. *Neuroscience Letters* **710**, 132933 (2019). https://doi.org:https://doi.org/10.1016/j.neulet.2017.06.052

158 Restelli, L. M. *et al.* Neuronal Mitochondrial Dysfunction Activates the Integrated Stress Response to Induce Fibroblast Growth Factor 21. *Cell Rep* **24**, 1407-1414 (2018). https://doi.org:10.1016/j.celrep.2018.07.023

159 Li, Q. *et al.* Correlations of Cerebrospinal Fluid/Plasma Fibroblast Growth Factor 21 Ratio with Metabolic Parameters in Chinese Individuals of Normal Weight. *Clin Lab* **62**, 893-899 (2016). https://doi.org:10.7754/clin.lab.2015.150926

160 Steinacker, P. *et al.* Chitotriosidase (CHIT1) is increased in microglia and macrophages in spinal cord of amyotrophic lateral sclerosis and cerebrospinal fluid levels correlate with disease severity and progression. *Journal of Neurology, Neurosurgery & Psychiatry* **89**, 239-247 (2018). https://doi.org:10.1136/jnnp-2017-317138

161 Varghese, A. M. *et al.* Chitotriosidase, a biomarker of amyotrophic lateral sclerosis, accentuates neurodegeneration in spinal motor neurons through neuroinflammation. *Journal of Neuroinflammation* **17**, 232 (2020). https://doi.org:10.1186/s12974-020-01909-y

162 Varghese, A. M. *et al.* Chitotriosidase - a putative biomarker for sporadic amyotrophic lateral sclerosis. *Clinical Proteomics* **10**, 19 (2013). https://doi.org:10.1186/1559-0275-10-19

163 Thompson, A. G. *et al.* Multicentre appraisal of amyotrophic lateral sclerosis biofluid biomarkers shows primacy of blood neurofilament light chain. *Brain Communications* **4** (2022). https://doi.org:10.1093/braincomms/fcac029

164 Chiot, A. *et al.* Modifying macrophages at the periphery has the capacity to change microglial reactivity and to extend ALS survival. *Nature Neuroscience* **23**, 1339-1351 (2020). https://doi.org:10.1038/s41593-020-00718-z