Improving the genetic diagnosis of Mendelian disorders using RNA-sequencing

David Zhang

Supervisor(s): Dr. Mina Ryten
Prof. John Hardy

UCL Great Ormond Street Institute of Child Health

Thesis submitted in fulfilment of the degree of Doctor of Philosophy

March 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. The work presented in this thesis is my own. Where the outcome of work is derived from collaboration and other sources, I confirm this has been specified in the text and Acknowledgements.

David Zhang
March 2022

Publications arising from this thesis:

- **Zhang D**, Reynolds RH, Ruiz SG, Gustavsson EK, Sethi S., Aguti S, Barbosa IA, Collier JJ, Houlden H, McFarland R, Muntoni F, Oláhová M, Poulton J, Simpson M, Pitceathly RDS, Taylor RW, Zhou H, Deshpande C, Botia JA, Collado-Torres L, Ryten M. Detection of pathogenic splicing events from RNA-sequencing data using dasper. *preprint - in submission*. 2021.

- **Zhang D**, Guelfi S, Ruiz SG, Costa B, Reynolds RH, D'Sa K, Liu W, Courtin T, Peterson A, Jaffe AE, Hardy J, Botia JA, Collado-Torres L and Ryten M. Incomplete annotation of disease-associated genes is limiting our understanding of Mendelian and complex neurogenetic disorders. *Science advances*. 2020.

Publications not directly related to this thesis:

- Sethi S, **Zhang D**, Guelfi S, Chen Z, Garcia-Ruiz S, Ryten M, Saini H, Botia JA. Leveraging omic features with F3UTER enables identification of unannotated 3'UTRs for synaptic genes. *Nature Communications - In submission*. 2021.

- Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, Imada EL, **Zhang D**, Joseph L, Leek JT, Jaffe AE, Nellore A, Collado-Torres L, Hansen KD, Langmead B. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biology - In press*. 2021.

- Fairbrother-Browne A, Ali AT, Reynolds RH, Garcia-Ruiz S, **Zhang D**, Chen Z, Ryten M, Hodgkinson A. Mitochondrial-nuclear cross-talk in the human brain is modulated by cell type and perturbed in neurodegenerative disease. *Communications Biology*. 2021.

- Collier J, Guissart C, Oláhová M, Sasorith S, Piron-Prunier F, Suom Fi, **Zhang D**, Martinez-Lopez N, Leboucq N, Bahr A, Azzarello-Burri S, Reich S, Schöls L, Polvikoski TM, Meyer P, Larrieu L, Schaefer AM, Alsaif HS, Alyamani S, Zuchner S, Barbosa IA, Deshpande C, Pyle A, Rauch A, Synofzik M, Alkuraya FS, Rivier F, Ryten M, McFarland R, Delahodde A, McWilliams TG, Koenig M, and Taylor RW. Developmental Consequences of Defective ATG7-Mediated Autophagy in Humans. *The New England Journal of Medicine*. 2021.

- Wilks C, Ahmed O, Baker DN, **Zhang D**, Collado-Torres L, Langmead B. Megadepth: efficient coverage quantification for BigWigs and BAMs. *Bioinformatics*. 2021.

- Kia DA, **Zhang D**, Guelfi S, Manzoni C, Hubbard L, United Kingdom Brain Expression Consortium (UKBEC), International Parkinson's Disease Genomics Consortium (IPDGC), Reynolds RH, Botía JA, Ryten M, Ferrari R, Lewis PA, Williams N, Trabzuni D, Hardy J, Wood NW. Integration of eQTL and Parkinson's disease GWAS data implicates 11 disease genes. *Jama Neurology*. 2021.

- Chen Z, **Zhang D**, Reynolds RH, Gustavsson EK, Garcia-Ruiz S, D'Sa K, Fairbrother-Brown A, Vandrovcova J, International Parkinson's Disease Genomics Consortium (IPDGC), Hardy J, Houlden H, Gagliano SA, Botiá J, Ryten M. Human-lineage-specific genomic elements: relevance to neurodegenerative disease and APOE transcript usage. *Nature Communications*. 2021.

- Guelfi S, D'Sa K, Botía J, Vandrovcova J, Reynolds RH, **Zhang D**, Trabzuni D, Collado-Torres L, Thomason A, Leyton PQ, Gagliano SA, Nalls MA, UK Brain Expression Consortium, Small KS, Smith C, Ramasamy A, Hardy J, Weale ME, Ryten M. Regulatory sites for known and novel splicing in human basal ganglia are enriched for disease-relevant information. *Nature Communications*. 2020.

- Salih DA, Bayram S, Guelfi S, Reynolds RH, Shoai M, Ryten M, Brenton JW, **Zhang D**, Matarin M, Botia JA, Shah R, Brookes KJ, Guetta-Baranes T, Morgan K, Bellou E, Cummings DM, Escott-Price V, Hardy J. Genetic variability in response to AB deposition influences Alzheimer's risk. *Brain Communications*. 2019.

- Holder-Espinasse M, Jamsheer A, Escande F, Andrieux J, Petit F, Sowinska-Seidler A, Socha M, Jakubiuk-Tomaszuk A, Gerard M, Mathieu-Dramard M, Cormier-Daire V, Verloes A, Toutain A, Plessis G, Jonveaux P, Baumann C, David A, Farra C, Colin E, Jacquemont S, Rossi A, Mansour S, Ghali N, Moncla A, Lahiri N, Hurst J, Pollina E, Patch C, Ahn JW, Valat AS, Mezel A, Bourgeot P, **Zhang D**, Manouvrier-Hanu S. Duplication of 10q24 locus: broadening the clinical and radiological spectrum. *Eur J Hum Genet*. 2019.

- Jabbari E, John W, Tan MMX, Maryam S, Pittman A, Ferrari R, Mok KY, **Zhang D**, Reynolds RH, de Silva R, Grimm MJ, Respondek G, Muller U, Al-Sarraj S, Gentleman SM, Lees AJ, Warner TT, Hardy J, Revesz T, Hoglinger GU, Holton JL, Ryten M and Morris HR. Variation at the TRIM11 locus modifies Progressive Supranuclear Palsy phenotype. *Annals of Neurology*. 2018.

Acknowledgements

This PhD has been an at times, difficult and above all, enriching journey. I would like to express my gratitude to all those who supported me along the way and without whom, I could not have reached this destination.

Foremost, I would like to thank the clinicians and patients who have made this research possible.

Most of all, I must thank Prof. Mina Ryten who has always been a beacon of patience, empathy, enthusiasm and knowledge. Her guidance and encouragement have been not only invaluable for the PhD, but also my development into the person I am today - I could not have wished for a better supervisor. I am also extremely grateful for the mentorship of Dr. Leonardo Collado-Torres, which was always given so generously. It was through his advice that I fully realised my confidence and passion for programming, a lifelong reward.

My parents, Lisheng Zhang and Keyu Fu, have been a pillar of unconditional support throughout my life and this PhD has been no different. I would not be here without them and I believe that this thesis serves as much their achievement as it is mine. Celestia Godbehere has been my rock that has kept my sanity from being washed away during the rising tides of the latter half of my PhD. I am extremely lucky to have her in my life and for her care and support, I will be forever grateful.

Many of the lessons learnt and the fun times had during my PhD came from interacting with my wonderful colleagues. These experiences are too many even for a bioinformatician to `list()`, but I am grateful for them all. I believe I have learnt the most and still have the most to learn from Regina Reynolds; your constant kindness and discipline is an inspiration. In Sebastian Guelfi, I have found a mentor, a colleague and a friend; I truly appreciate the countless (mumbo jumbo) highs and also the few not-so-highs we have shared together. A warm thanks goes to my colleagues and friends whom I have worked hard and played hard with; Aine Fairbrother-Browne,

Beatrice Costa, Dr. Emil Gustavsson, Dr. Siddharth Sethi and Sonia Garcia-Ruiz. Together, you have all made this journey a profoundly enjoyable one.

Abstract

Providing patients with Mendelian disorders a genetic diagnosis improves the management of symptoms, informs genetic counselling and provides opportunities for therapeutic intervention. The advent of next generation sequencing technologies have greatly improved our ability to identify gene-disease associations. Despite these advances, most patients still leave the clinic without a genetic diagnosis. Although whole genome sequencing can capture genome-wide genetic variation, accurate interpretation of these variants remains a major challenge. In this thesis, I develop and apply methods that use transcriptomics to improve variant interpretation and consequently, diagnostic yield.

Using publicly available RNA-seq data across 43 different human tissues, I improved the annotation of the majority of known, disease-causing genes. The detected novel exons were more depleted for genetic variation in humans than expected by chance, suggestive of their functional importance. In addition, a subset were shown to be potentially protein-coding. The novel annotation is publicly released through the resource, vizER, which enables the querying and visualisation of genes for evidence of their reannotation.

I developed the R/Bioconductor package, *dasper*, which integrates junction and coverage data from RNA-sequencing to improve the detection of aberrant splicing events. Benchmarking analysis demonstrated that dasper detects pathogenic splicing events more accurately than existing approaches, as well as harnesses both publicly available and in-house RNA-sequencing data effectively as controls. dasper is designed for diagnostics, providing a rank-based report of how aberrant each splicing event looks, as well as including visualization functions to facilitate interpretation.

RNA-sequencing was applied to fibroblasts derived from a cohort of patients with suspected mitochondrial disorders, who remained unsolved after whole exome sequencing. Using this approach, a genetic diagnosis was confirmed for 1 patient and candidate genes were discovered for a remaining third. Beyond diagnosis, the potential of RNA-sequencing to improve our understanding of disease pathogenesis was explored in three ways; deriving the mechanism of splicing disruptions, detection of perturbed pathways downstream of the pathogenic variant and elucidating genetic modifiers that influence phenotypic variability.

Impact Statement

Within this thesis, I contribute insights, tools and resources that will facilitate the application of RNA-seq for diagnostics for researchers and diagnostic laboratories, as well as increase diagnostic yield for the benefit of patients with Mendelian disorders.

The analyses within this thesis contribute broad insights for the field of diagnostic RNA-seq. In chapter 1, my work describes how diagnostic rates remain hindered by incomplete annotation, and the disproportionate impact this is likely to have on neurogenetic disorders. These results can be helpful for determining when unsuccessful diagnoses are likely to be attributed to incomplete annotation, as well as offer a potential solution through the related resource, *vizER*. In chapter 2, the development of *dasper* highlights the possibility of using publicly available RNA-seq data as controls for diagnostic tools. If put into practice, this is likely to greatly improve the accessibility of RNA-seq in a diagnostic setting. In chapter 3, I describe how RNA-seq can be used to systematically characterise the consequence of aberrant splicing events. This informs the development of future methods which have the aim of automating the interpretation of aberrant splicing events using short-read RNA-seq; an area that remains underrepresented in existing tools, but holds considerable promise for the design of splice-modulating therapies.

The work in this thesis has also led to the development of two Bioconductor packages, *dasper* and *ODER*, that facilitate the application of RNA-seq by researchers. The output of these tools can be integrated into a diagnostic RNA-seq pipeline, with the goal of improving variant interpretation and, potentially, diagnostic yield. I have designed these tools to be robust and user-friendly by conforming to best practices of software development. For example, both packages are extensively tested across three operating systems (Windows, Mac, Linux) and have their documentation continuously deployed via a *pkgdown* website. Furthermore, the release of these tools on the Bioconductor platform improves their visibility and credibility. Together, these factors are likely to increase the user base of the tools and at the time of this thesis submission, *dasper* and *ODER* have collectively been downloaded over 1300 times.

In chapter 3, I apply diagnostic RNA-seq to a cohort of 32 patients with suspected mitochondrial disorders, diagnosing a single patient and finding candidate genes for a third. In the diagnosed patient, a successful genetic diagnosis of the patient's disorder will permit more accurate genetic counselling for the patient's parents, in particular for family planning. For the remaining third of patients, which had candidate genes identified, these are currently being followed up in ongoing investigations.

Overall, I anticipate that the work in this thesis will facilitate the wider adoption of diagnostic RNA-seq as well as potentially increase the diagnostic yield achievable through its application. Ultimately, it is my hope that this will contribute to improvements in the lives of patients with Mendelian disorders.

Table of contents

List of figures

List of tables

Abbreviations

ACMG	American College of Medical Genetics
ASE	Allele-specific expression
CDTS	Context dependent tolerance scores
CNS	Central nervous system
DJs	Downregulated junctions
ERs	Expressed regions
FPKM	Fragments per kilobase of transcript per million mapped reads
GTEx	Genotype-Tissue Expression Consortium
GWAS	Genome-wide association studies
HPO	Human Phenotype Ontology
iPSCs	Induced pluripotent stem cells
Mb	Megabases
MCC	Mean coverage cut-off
MRG	Mean region gap
NGS	Next-generation sequencing
OMIM	Online Mendelian Inheritance in Man
$RNA-seq$	RNA-sequencing
SNPs	Single nucleotide polymorphisms

TPM	Transcripts per kilobase per million mapped reads
UJs	Upregulated junctions
UTRs	Untranslated regions
VUSs	Variants of unknown significance
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Chapter 1

Introduction

1.1 Genetic diagnosis of Mendelian disorders

Mendelian disorders, collectively affecting more than 1 in 50 individuals, impose a considerable burden on healthcare systems worldwide (1). Accurate molecular diagnosis of Mendelian diseases improves the management of patient symptoms, informs genetic counselling and provides opportunities and preventative therapies (2). The advent of next-generation sequencing (NGS) has accelerated the identification of causative genes and variants associated with disease and as result, whole-exome sequencing (WES) and whole-genome sequencing (WGS) are increasingly incorporated into the diagnostic routine. WES offers a cost-effective, comprehensive approach to capture pathogenic variants within protein-coding regions and, depending on the disorder and the selection of patients, has been estimated to lead to a diagnosis rate of 30%–50% (3). Patients that are left without a molecular diagnosis after exome-sequencing likely remain unsolved as their pathogenic variant(s) reside in novel disease genes or are not readily detectable by WES. These include structural variants, repeat expansions, and those that lie within deep intronic regions of the genome (4). For such cases, WGS has been demonstrated as a promising approach to elucidate pathogenic variants missed by WES (5). Although WGS enables the detection of variants at a genome-wide scale, accurate interpretation of the functional consequence of the captured genetic variation remains a major challenge. In the context of establishing a molecular diagnosis, this leads to patients being left undiagnosed, since pathogenic mutations cannot be accurately distinguished from the other rare, potentially functional, yet benign variants present in any human genome (6). Importantly, the interpretation of non-coding variants is particularly difficult, due to the historical prioritisation and consequently, better understanding of protein-coding regions of the genome (7, 8). Furthermore, it is increasingly recognised that these non-coding regions confer important regulatory roles and consequently, non-coding variants have the capability to cause Mendelian disorders (8, 9). Overall, in large part due to the limitations in variant interpretation, the current rate of successful genetic diagnoses remains an estimated 25-50% (5, 10, 11). One possible solution lies in characterising the downstream functional consequences of variants, for instance, at the transcriptome, proteome or metabalome level (3). These approaches enable the discovery of pathogenic molecular products, which can be used to re-interpret variants and resolve variants of unknown significance (VUSs), leading to an increase in diagnostic yield.

1.2 Incomplete gene annotation hinders variant interpretation

Reference annotation databases are comprised of the genomic co-ordinates of all the genes, transcripts and their constituent exons in a given model organism. Current efforts to annotate the human genome principally originate from 4 organisations; RefSeq, GENCODE, Ensembl and AceView, which differ in the stringency of their annotation pipeline and consequently, the demographic of their user base (12–15). For example, the reference annotation derived from RefSeq is the most conservative and therefore, most commonly used within the diagnostic community. While these databases have now become essential resources within research and diagnostic analyses, there is evidence to suggest that they remain incomplete or inaccurate. Foremost, reference annotation databases are still consistently updated, with the addition of newly discovered genes as well as corrections to existing gene definitions (12). Secondly, there are discrepancies between the reference annotations generated by different sources (16, 17). Thirdly, analyses performed using large-scale RNA-sequencing (RNA-seq) datasets have revealed an abundance of transcription originating from intergenic or intronic (unannotated) regions (18). Finally, efforts using *de novo* transcript assembly methods or long-read RNA-seq have discovered a plethora of unannotated transcripts, especially for genes with complex splicing patterns (19, 20).

The interpretation of the functional consequence of a variant on a gene or transcript is fundamentally reliant on accurate and complete reference annotation (Figure 1.1) (21). This is exemplified by changes in variant interpretation arising from the usage of different reference annotation databases. For instance, Frankish et al. found that variant annotation was substantially different when using GENCODE versus RefSeq derived reference annotation. In addition, these discrepancies were concentrated on variants that fell into non-coding transcripts or untranslated regions (UTRs) (17). In the context of genetic diagnoses, this suggests inaccurate or incomplete gene annotation can lead to variants being falsely de-prioritised as benign or assigned as VUSs. Together, it is likely that improvements to reference annotation will lead to more accurate interpretation of variants and therefore, an increase in the rate of genetic diagnoses.

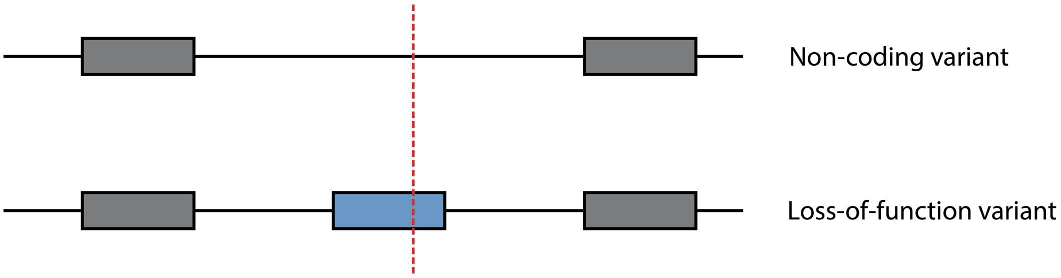


Figure 1.1 Variant interpretation relies on reference annotation. A simple schematic to illustrate how the interpretation of a non-coding variant (red, dashed line) could shift depending on genome annotation. Annotated, known exons are represented by grey boxes, whilst the blue box marks a discovered novel exon. Importantly, the discovery of novel exons could permit the prediction of the consequence of any overlapping variants on protein sequence and function.

1.3 Detection of aberrant events using RNA-seq

In the past decade, RNA-seq has become the gold-standard approach for measuring RNA abundance and transcript diversity. More recently, it has also been shown to be a promising diagnostic tool for cases that remain unsolved after WES or WGS (22). RNA-seq has been applied in diagnostic pipelines to obtain a transcriptome-wide readout of gene expression, splicing levels and allele-specific expression (ASE) in patient samples (23, 24). By comparing these metrics to a set of unaffected controls, this enables the detection of aberrant RNA-level products, namely aberrantly expressed genes, aberrantly spliced transcripts and allele-specific expression events (Figure 1.2). Such aberrant events serve as functional data, which provide evidence for the downstream consequences of variants with the patient sample. In this way, the RNA-seq derived aberrant events complement the DNA-seq derived variants, enabling more accurate variant interpretation. The re-interpretation of variants using RNA-seq can resolve variants of unknown significance (VUSs) and/or narrow down a list of candidate variants to an actionable size permitting further functional investigation (25). Together, using this approach, RNA-seq has been leveraged to detect aberrant events, re-prioritise variants and lead to assignment of pathogenicity.

To date, several studies have applied RNA-seq for diagnostics, achieving an improvement in diagnostic yield between 5-35% for cases unsolved after WES or WGS (26, 27). Earlier studies have focused on exome-negative patient cohorts skeletal muscle disorders or mitochondrial disease, however the more recently, diagnostic RNA-seq has also been applied to cohort with a large variety of disorders (23, 24, 27–29). In order to systematically assess the value of integrating RNA-seq into the diagnostic pipeline, a recent study investigated the diagnostic yield of using WES, WGS and RNA-seq for 113 rare disease patients spanning a wide spectrum of clinical indications (25). In total, a genetic diagnoses was achieved for 38% of the 113 patients; the diagnoses rate using WES/WGS alone was 31%, whilst RNA-seq evidence was required to establish the additional 7% of diagnoses. Overall, these studies have established RNA-seq as a valuable diagnostic tool that can be effective across a wide range of disorders.

A major challenge of using diagnostic RNA-seq arises from the differences in expression and splicing across human tissues, coupled with the fact that disease relevant tissues are not always accessible; this is commonly referred to as the proxy tissue problem (26). In this way, the proxy tissues sampled from patients may not express the disease-associated RNA-level events to a detectable level, thereby limiting the diagnostic yield from RNA-seq. Despite this limitation, recent studies highlight that pathogenic events can still be successfully detected in tissues that are typically unaffected by disease. For example, Fresard et al. use RNA-seq from whole blood samples to successfully diagnose 7.5% of a cohort of 94 rare disease patients with clinical diagnoses spanning 16 diverse disease categories, including neurological disorders (27). In order to quantify the the extent to which the proxy tissue problem will limit RNA-seq for diagnostics, a recent study has assessed the ability of clinical accessible tissues (CATs) to capture the transcriptomic complexity of the remaining human tissues. With this approach, they estimate the majority (60%) of genes had splicing adequately replicated in at least 1 CAT (30). Furthermore, they release their findings as a database, *MAJIQ-CAT*, upon which the known disease genes can be queried to gauge the suitability of a patient or disorder for the diagnostic RNA-seq approach. In recent years, new technologies have emerged that allow the differentiation of CATs into disease relevant tissues; these represent a potential solution to the proxy tissue problem. For example, Gonorazky et al. and Bronstein et al. demonstrate that differentiation of fibroblasts or induced pluripotent stem cells (iPSCs) to disease tissues enabled the detection of pathogenic events that would otherwise have been missed in the original tissue (28, 29). Overall, these studies demonstrate that RNA-seq has the potential to provide diagnostic value, even in situations where the disease relevant tissue is inaccessible. Furthermore, it is forseeable that resources such as *MAJIQ-CAT* and technologies such as IPSCs will reduce the impact of the proxy tissue problem and thereby improve the diagnostic value of RNA-seq in the future.

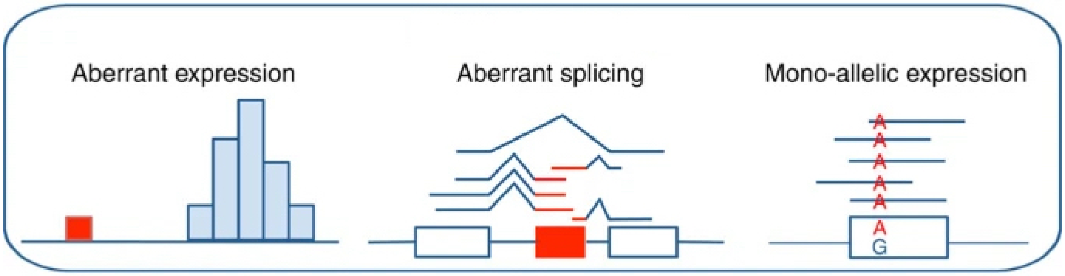


Figure 1.2 RNA-sequencing for diagnostics workflow. Adapted from Kremer et al. 2017 (24). There are three major strategies employed to facilitate genetic diagnosis of Mendelian disease patients using RNA-seq: detection of aberrantly expressed genes, aberrantly spliced events and mono-allelic expression of the alternative allele. Such aberrant events can be used to complement DNA-seq sequencing and improve variant interpretation, leading to potential increase in diagnostic yield.

1.4 Tools for the detection of aberrant events from RNA-seq

As the number of studies applying diagnostic RNA-seq have increased, so too has the number bioinformatic tools developed for its application. Currently, to my knowledge, there are 4 diagnostic tools that are released for the detection of aberrant events from RNA-seq, which can be categorised by the type of aberrant event which they aim to discovery; aberrant expression, aberrant splicing and ASE events (Table 1.1). Traditional differential expression or splicing tools such as *DESeq2* and *leafcutter* have been designed to compare between groups of samples, typically across treatment versus control, often with a few replicates for each sample (31, 32). These approaches do not transfer well to Mendelian disease diagnostics, where there are often no replicates available and there is typically an experimental design where a single patient is compared to a set of controls (1-vs-all). In addition, one is often interested in identifying an outlier event in this single patient, rather than a event with a subtle fold change between groups. For a result, all diagnostic RNA-seq tools developed to date have employed a 1-vs-all framework, which aims to detect outlier/aberrant events by comparing each patient with a set of controls (33–36).

1.5 The value of splicing information in a clinical setting

Alternative splicing is the complex, tightly regulated process by which introns are excised from pre-mRNA. Splicing is an essential cellular process used to generate transcriptomic and functional

	Tool	Type	DOI
1	OUTRIDER	Expression	doi.org/10.1016/j.ajhg.2018.10.025
2	FRASER	Splicing	doi.org/10.1038/s41467-020-20573-7
3	LeafCutterMD	Splicing	doi.org/10.1093/bioinformatics/btaa259
4	ANEVADOT	Allele-specific expression	doi.org/10.1126/science.aay0256

Table 1.1 Bioinformatic tools for diagnostic RNA-seq. The table highlights the current available tools that are designed to facilitate diagnostic RNA-seq through the detection of aberrant events from RNA-seq.

complexity in higher eukaryotes (37). The spliceosome complex is comprised of hundreds of proteins and small nuclear RNAs, which function in concert to form the machinery required for splicing (38). The fidelity of splicing also depends on interactions between trans-acting factors (proteins and ribonucleoproteins) and cis-acting pre-mRNA sequence motifs, which regulate splicing through facilitating the binding of splicing factors (39). The complexity of splicing is crucial for generating transcript and phenotypic diversity, but also increases the vulnerability of splicing to perturbations. Consequently, aberrant splicing has been shown to be a key cause of Mendelian disorders, with an estimated one third of pathogenic mutations impacting splicing (40, 41). Notably, pathogenic variants that disrupt splicing most commonly fall within the core consensus motifs (5' splice site, 3' splice site, and branch point) or within non-coding regulatory elements.

In the past decade, RNA-seq has become the principal approach used to obtain a transcriptome-wide profile of alternative splicing. From a short-read RNA-seq experiment, information regarding splicing can be obtained in two main ways; measures of junction counts and coverage (Figure 1.3) (42). Junctions reads are defined as those reads which map with a gapped alignment to the genome, with the gap representing the removal of an intron via the mechanism of splicing. The number of reads across a given junction can be used to quantify the relative expression of a splicing event. On the other hand, coverage is defined as the number of reads that align to each base within the genome. Coverage across exons can be used to derive the relative usage of exons, which in turn can inform the number of times an intron is excised. Overall, RNA-seq can provide a transcriptome-wide profile of the splicing within a given sample through measures of junction counts and coverage.

The value of splicing information obtained through RNA-seq is particularly valuable in a clinical context for two reasons. Firstly, variants distributed in the non-coding regions of the genome disproportionately affect splicing, often through disruptions to intronic splicing enhancers, silencers or recognition sequences (8). Therefore, in a diagnostic context, detection of aberrant splicing events using RNA-seq can be useful for the re-interpretation of the non-coding variation. Consistent with this, the majority of molecular diagnoses achieved using RNA-seq to date have involved information regarding disruptions in splicing (23, 24, 27). Secondly, splicing information has the potential to inform the development of splice-modulating treatments such as antisense oligonucleotides (ASOs) (43). For example, Kim et al. applied RNA-seq characterise a pathogenic splicing event and as a results, design and target a personalised ASO for a patient with Battens disease (44). Together, RNA-seq provides a platform to measure transcriptome-wide levels of alternative splicing in a patient, which can be used to improve diagnostic yield through the re-interpretation of non-coding variation, as well as inform the development of splice-modulating therapies.

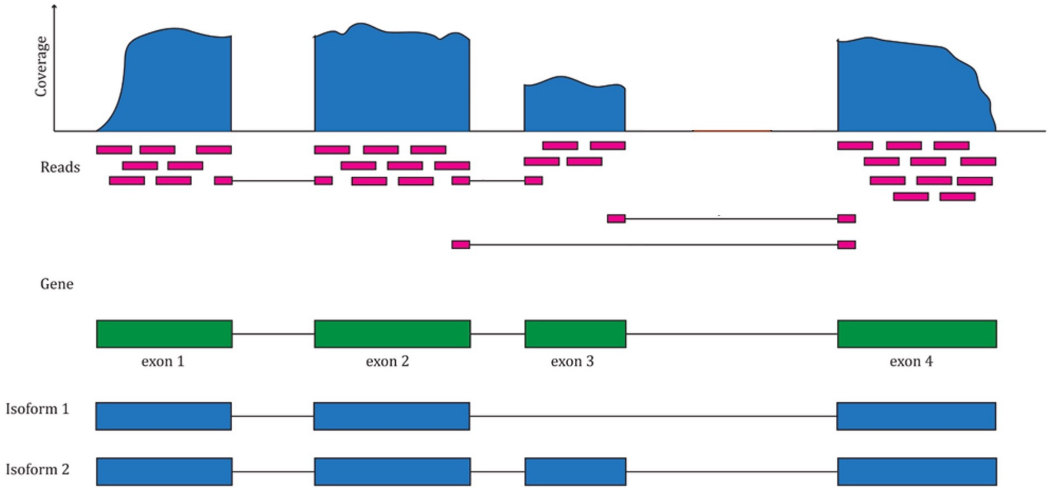


Figure 1.3 Detection of splicing from RNA-sequencing. Adapted from Collado-Torres et al. 2017 (42). RNA-seq reads are represented by the pink boxes. Normal reads are those that fall completely within exonic regions, which can be used to inform the coverage and therefore, the differential usage of exons. Junction reads are those that align to the boundary of two exons with a gap within the middle that marks the excision of an intron. Together, junctions and coverage can be used to capture, measure and quantify splicing events from RNA-seq.

1.6 Objectives of this thesis

The overarching objective of this thesis is to demonstrate and improve the value of RNA-seq as a diagnostic tool for Mendelian disorders. I address this goal in 2 ways; 1. developing and releasing resources and pipelines that facilitate the use of RNA-seq in a diagnostic setting and 2. apply RNA-seq to diagnose a cohort of patients that remain unsolved after WES.

In chapter 2, I leverage RNA-seq across multiple human tissues to improve the annotation for the majority of disease genes. The annotation I generate is publicly released via a web interface *vizER*, on which individual genes can be queried for novel exons. In addition, the pipeline for the generation of this novel annotation is made available as a Bioconductor package, *ODER*, which facilitates the re-application of the annotation pipeline to additional datasets.

In chapter 3, I develop and publish the Bioconductor package, *dasper*, which uses and junctions and coverage to detect aberrant splicing events from RNA-seq data. *dasper* is tailored for use in a diagnostic setting; it applies a 1-vs-all experimental framework to detect outlier splicing events and has in-built visualisation functionality to facilitate the interpretation of detected splicing events.

In chapter 4, I apply RNA-seq to diagnose a set of unsolved patients with suspected mitochondrial disorders. Using this approach, I diagnose 1 patient and obtain improved candidate gene resolution for a remaining third of patients. Candidate genes are currently being followed up for further investigation by the clinicians. Furthermore, I touch upon the use of RNA-seq for utility beyond diagnoses, namely the characterisation of the consequence of splicing events and improvements to disease understanding.

Together, it is my hope, that the work within this thesis expands upon and facilitates the use of RNA-seq as a tool for diagnostics and beyond.

Chapter 2

Improving gene annotation of disease-causing genes using RNA-sequencing data

2.1 Introduction

Genetic and transcriptomic studies are fundamentally reliant on accurate and complete human gene annotation. Amongst other analyses, this is required for the quantification of expression or splicing from RNA-sequencing experiments, interpretation of significant genome-wide association studies (GWAS) signals and variant interpretation from genetic tests. As the understanding of transcriptomic complexity improves, it is apparent that existing gene annotation principally originating from 4 sources (RefSeq, GENCODE, Ensembl, AceView) remains incomplete (12–15). Comparison of these different existing gene annotation databases reveals that over 17,000 Ensembl genes fall into intronic or intergenic regions according to the AceView database and the choice of reference annotation greatly influences the performance of variant interpretation software, such as VEP and ANNOVAR (16, 18). This evidence suggests that incomplete annotation may cause pathogenic variants to be overlooked within exonic regions that are yet to be annotated.

Despite accumulating evidence that the map of the human transcriptome remains incomplete, it is not yet fully understood which tissues and consequently diseases are most affected. The extent to which this poses an issue is unlikely to be equal across all types of tissues or cells. In particular, the fact that the human brain harbours longer transcripts, higher transcript diversity and higher cellular heterogeneity than other tissues might be expected to make identifying all transcripts from this tissue more challenging (45). Moreover, the difficulties of accessing brain tissue and dependence on post-mortem tissue may also limit the quantity of high quality, brain-specific data inputted into gene annotation pipelines to date. In fact, several analyses of bulk RNA-sequencing data derived from human brain tissues have discovered transcription originating from intronic or intergenic regions (henceforth termed novel) (46, 47). For example, Jaffe and colleagues found that as much as 41% of transcription in the human frontal cortex was novel (18). In combination, these factors lead to specific challenges in fully capturing the transcriptome of the human brain and suggest that improvements to gene annotation may have a disproportionate impact on the understanding of neurological diseases.

In this study, I address this issue by leveraging transcriptomic data available through the Genotype-Tissue Expression Consortium (GTEx) to identify novel exons of known disease genes. Distinct from existing de novo assembly approaches, such as that implemented by Pertea and colleagues leading to the development of the CHESS database, my analytical approach was focused on the detection of novel exons amongst known genes rather than the assembly of novel transcripts (19). This conservative approach was adopted because of the well-recognised challenges in accurately calling novel transcripts from short read sequencing data and because the major aim of this study was to improve the annotation of genes already known to cause Mendelian disorders (48, 49). With this in mind, I defined transcription in an tissue-specific, annotation-agnostic manner using RNA-sequencing data from 13 regions of the human central nervous system (CNS) and a further 28 non-brain tissues. I found that novel transcription although widespread across all tissues, is most prevalent in human brain. I provide evidence to suggest that the novel exons I discover are likely to be functionally important on the basis of their tissue specific expression, the significant depletion of genetic variation within humans and their protein-coding potential. Finally, by combining novel transcription with junction read data, defined as reads that have a gapped alignment to the genome, I link these regions to known genes, focussing on those associated with Mendelian disorders. Overall, I improve the annotation of 13,429 genes, encompassing 1831 (63%) OMIM genes. I release this data via an online platform *vizER* as well as the method for its generation as a Bioconductor package, *ODER*. *vizER* allows individual genes to be queried and visualised for re-annotation as well as the download of all novel annotations discovered. It is my hope that this resource will facilitate improvements to the genetic diagnosis of Mendelian disorders.

2.2 Methods

2.2.1 OMIM data

Phenotype relationships and clinical synopses of all Online Mendelian Inheritance in Man (OMIM) genes were downloaded using http://api.omim.org on the 29th of May 2018 (50). OMIM genes were filtered to exclude provisional, non-disease and susceptibility phenotypes retaining 2,898 unique genes that were confidently associated to 4,034 Mendelian diseases. Phenotypic abnormality groups were linked to corresponding affected GTEx tissues through manual inspection of the Human Phenotype Ontology (HPO) terms within each group by a medical specialist (51).

2.2.2 GTEx data

Data download and wrangling of the GTEx data was performed by Sebastian Guelfi. RNA-seq data in base-level coverage format for 7,595 samples originating from 41 different GTEx tissues was downloaded using the R package *recount* version 1.4.6 (52). Cell lines, sex-specific tissues and tissues with 10 samples or below were removed. Samples with large chromosomal deletions and duplications or large CNVs previously associated with disease were filtered out (smafrze = "USE ME"). Coverage for all remaining samples was normalised to a target library size of 40 million 100bp reads using the area under coverage value provided by *recount2*. For each tissue, base-level coverage was averaged across all samples to calculate the mean base-level coverage. GTEx junction read data, defined as reads with a non-contiguous gapped alignment to the genome, was downloaded using the *recount2* resource and filtered to include only junction reads detected in at least 5% of samples for a given tissue and those that had available donor and acceptor splice sequences.

2.2.3 Optimising the detection of transcription

This method was concieved with the help of Sebastian Guelfi, a postdoctoral research associate within the Ryten lab at the time this analysis was conducted. Transcription was detected across 41 GTEx tissues using the package derfinder version 1.14.0 (53). The mean coverage cut-off (MCC), defined as the number of reads supporting each base above which bases were considered to be transcribed, and max region gap (MRG), defined as the maximum number of bases between expressed regions (ERs) below which adjacent ERs will be merged, were optimised.

Optimisation was performed using 156,674 non-overlapping exons (defined by Ensembl v92) as the gold standard (12). Exon biotypes of all Ensembl v92 exons were compared to this set of non-overlapping exons to ensure the analysis was not preferentially optimising for one particular biotype (Figure 2.1). Non-overlapping exons were selected as these definitions would be least likely to be influenced by ambiguous reads. For each tissue, I generated ERs using mean coverage cut-offs increasing from 1 to 10 in steps of 0.2 (46 cut-offs) and max gaps increasing from 0 to 100 in steps of 10 (11 max region gaps) to produce a total of 506 unique transcriptomes. For each set of ERs, I found all ERs that intersected with non-overlapping exons, then calculated the exon delta by summing the absolute difference between the start/stop positions of each ER and the overlapping exon (Figure 2.3a). Situations in which a single ER overlapped with multiple exons were removed to avoid assigning the ER to an incorrect exon when calculating downstream optimisation metrics. For each tissue, I selected the mean coverage cut-off and max region gap, which minimised the difference between ER and "gold standard" exon definitions (median exon delta) and maximised the number of ERs that precisely matched the boundaries of exons (number of ERs with an exon delta equal to 0). All ERs that were <3bp in width were removed as these were below the minimum size of a microexon (54).

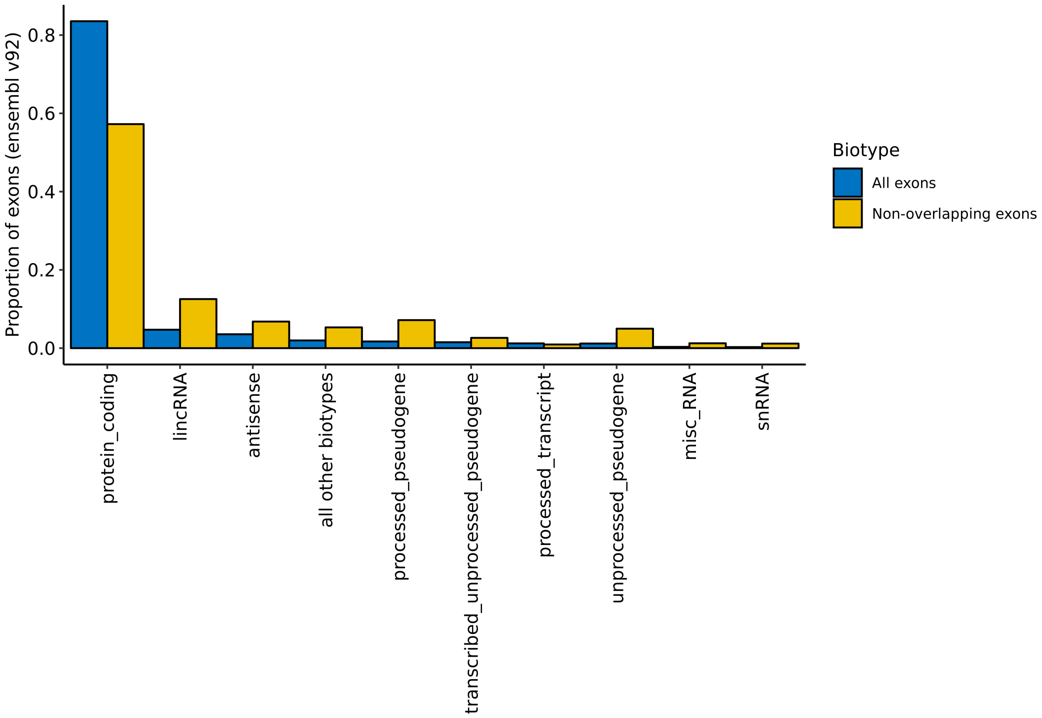


Figure 2.1 Proportion of exons that fall into different gene biotypes. Comparison of the proportion of exons that are classified within the different gene biotypes between all exons from Ensembl v92 and the non-overlapping set of exons used to optimise the detection of transcription.

2.2.4 Calculating the transcriptome size per annotation feature

ERs were classified with respect to the annotation feature (exon, intron, intergenic) with which they overlapped. A minimum of 1bp overlap was required for an ER to be categorised as belonging to a given annotation feature. ERs overlapping multiple annotation features were labelled with a combination of each. This generated 6 distinct categories – “exon”, “exon, intron”, “exon, intergenic”, “exon, intergenic, intron”, “intergenic” and “intron” (Figure 2.2a). ERs classified as “exon, intergenic, intron” were removed from all downstream analysis as these formed only 0.54% of all ERs and were presumed to be technical artefacts generated from regions of dense, overlapping gene expression. For each tissue, the length of all ERs within each annotation feature was summed generating the total Mb of ERs per annotation feature. Normalised variance of exonic, intronic and intergenic ERs was calculated by dividing the standard deviation of the total Mb of ERs across tissues by the mean total Mb of ERs for each annotation feature. To compare between brain and non-brain tissues, the total Mb of intronic and intergenic ERs were first summed together to generate an overall measure of novel transcription abundance across brain and non-brain tissues, then a two-sided Wilcoxon rank sum test was applied.

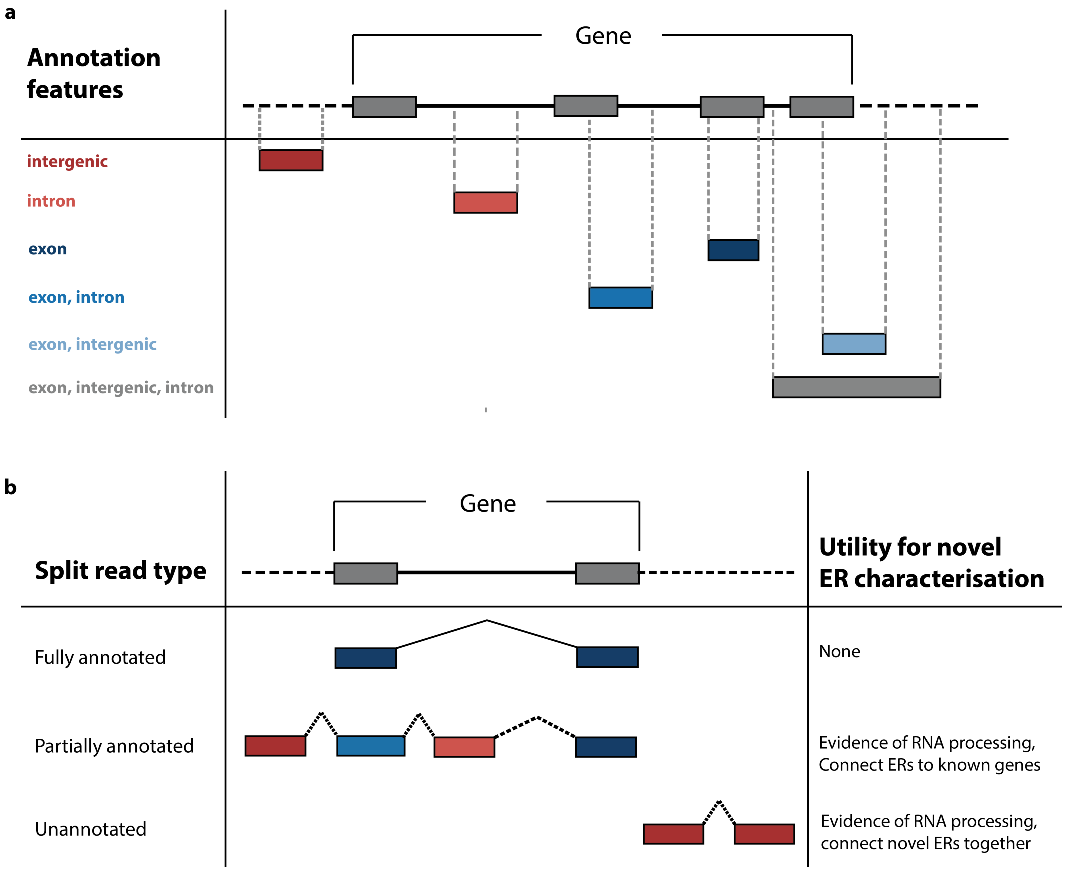


Figure 2.2 Characterising ERs using Ensembl annotation features and split reads. a) Illustration of the ER categorisation dependent on overlap with existing gene annotation. ERs in red are considered novel transcription. Blue ERs are those that overlap existing exons and are considered part of existing annotation. Grey ERs were uninformative and likely an artefact generated from genomic regions with high amounts of noise, pre-mRNA or overlapping genes, therefore were removed from all downstream analysis. b) Diagram showing the use of split reads (reads with a gapped alignment to the genome) to characterise novel ERs. Split reads were classified as annotated, partially annotated or unannotated dependent on whether the acceptor or donor sites both overlapped, only 1 of the acceptor or donor sites overlapped or neither overlapped known Ensembl v92 exon boundaries respectively. Partially annotated split reads were used to connect novel ERs to known genes. Partially annotated and unannotated split reads were used to provide evidence of RNA processing for novel ERs.

2.2.5 Annotating ERs with junction read data

Intronic and intergenic ERs were connected to known genes using reads, which I term junction reads, with a gapped alignment to the genome, presumed to be reads spanning exon-exon junctions (Figure 2.2b). Such exon-exon junctions are defined as non-contiguous reads which fall on the boundary between two exons of the same mRNA molecule, therefore when aligned to the genome these reads have a break in the middle indicating the splicing out of an intron. Junction read data was categorised into three groups: annotated junction reads, with both ends falling within known exons; partially annotated junction reads, with only one end falling within a known exon; and unannotated junction reads, with both ends within intron or intergenic regions.

In this way, intron and intergenic ERs that overlapped with partially annotated junction reads were connected to known genes.

2.2.6 Validation of detected transcription

Transcription was validated across different versions of Ensembl and within an independent dataset. ERs that overlapped purely intronic or intergenic regions according to Ensembl v87, but fell within exons according to v92, were counted as novel transcription that was validated in later versions of Ensembl. Furthermore, ERs overlapping exonic regions in Ensembl v87 now classified as intronic or intergenic in v92 were measured to control for expected corrections in gene definitions. To assess whether the total Kb of validated novel ERs entering v92 annotation was greater than what would be expected by chance, I generated 10,000 random sets of length-matched regions for each tissue that were intronic or intergenic with respect to Ensembl. Using a one sample Wilcoxon test, I compared the total Kb of intronic and intergenic ERs entering annotation to the total Kb distribution of the randomised intronic and intergenic regions, respectively.

Validation within an independent dataset was performed using RNA-seq coverage data from 49 control frontal cortex (BA9) samples originally reported by Labadorf and colleagues (2015) and available via the recount R package version 1.4.6 (52, 55). ERs derived from the GTEx frontal cortex (BA9) data were re-quantified using this independent frontal cortex dataset and those that had a mean coverage of at least 1.4 (the optimised MCC for the GTEx frontal cortex data), were counted as novel transcription that was validated.

2.2.7 Analysing the conservation and constraint of novel ERs

Conservation scores in the form of phastCons7 (derived from genome-wide alignments of 7 mammalian species) were downloaded from UCSC (56). Constraint scores generated from the genome-wide alignment of 7,794 unrelated human genomes were downloaded as context dependent tolerance scores (CDTS) (57). The raw phastCons7 and CDTS were in bins of 1bp and 10bp, respectively, therefore when annotating the corresponding positions of ERs, I aggregated each score as a mean across the entire genomic region of interest. To account for missing CDTS values, I calculated the coverage of each ER by dividing the number of bases annotated by the CDTS by the total length of the ER. For all downstream analysis, I filtered out ERs for which CDTS coverage was less than 80%.

To assess whether our novel ERs were more constrained or conserved than by expected by chance, I compared the phastCons7 and CDTS of novel ERs to 10,000 randomised length-matched sets of intronic and intergenic ERs for each tissue. For each of the 10,000 iterations, I first selected a random intronic or intergenic region that was larger than the respective ER, then selected a random segment along the randomised region which matched the length of the corresponding ER. The randomised regions were annotated with constraint scores and CDTS using the aforementioned method. The mean CDTS and phastCons7 of the novel ERs (split by annotation feature) were compared to the corresponding distribution of CDTS and phastCons7 of the randomised regions using a one sample, two-tailed t-test. For easier interpretation when plotting, CDTS scores have been converted to their opposite sign, therefore for both phastCons and CDTS, the higher the value the greater the magnitude of conservation or constraint.

2.2.8 Checking ER protein-coding potential

Intronic and intergenic ERs that were intersected by 2 junction reads were extracted. The junction reads were used to determine the precise boundaries of the ER. The DNA sequence corresponding to the ER genetic co-ordinates was extracted from the genome build hg38. Since the translation frame was ambiguous without knowledge of the other exons that are part of the transcript that included the novel ER, I converted the DNA sequence to amino acid sequence for all three possible frames starting from the first, second or third base. Any ER that had at least 1 frame that did not include a stop codon was considered to be potentially protein coding.

2.2.9 Gene properties influencing re-annotation

All Ensembl v92 genes were marked with a 1 or a 0 depending on whether I detected a re-annotation for that gene in the form of an ER connected to the gene using a junction read, with 1 representing a detected re-annotation event. Details of gene length, biotype, transcript count and whether the gene overlapped another gene were retrieved from the Ensembl v92 database. Brain-specificity was assigned using the Finucane dataset and selecting the top 10% of brain-specific genes when compared to non-brain tissues (58). Mean gene TPM was calculated by downloading tissue-specific TPM values from the GTEx portal and summarised by calculating the mean across all tissues. The list of OMIM genes was used to assign whether a gene was known to cause disease or not. I used a logistic regression to test whether different gene properties significantly influenced the variability of re-annotation.

2.2.10 Sanger sequencing of novel junctions

Primer design and sanger sequencing was performed with the help of other members of the Ryten lab; Regina Reynolds and Beatrice Costa. Commercially purchased (Takara) frontal cortex and cerebellum RNA samples, isolated from individuals of European descent, were used for validation of novel junctions detected in *SNCA* and *ERLIN1* respectively. Tissues were chosen to match the tissue in which the re-annotation for each gene was detected. Reverse transcription was performed using 1ug of RNA from each tissue, then converted to cDNA using the High-Capacity cDNA Reverse Transcription Kit with RNase Inhibitor (Applied Biosystems) and random primers as per manufacturer's instructions. Primers were designed to span predicted exon-exon junctions using Primer-BLAST (NCBI) and ordered from Sigma. PCR was performed using FastStart PCR Master (Roche) and enzymatic clean-up of PCR products was performed using Exonuclease I (Thermo Scientific) and FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific). Sanger sequencing was performed using the BigDye terminator kit (Applied Biosystems) and sequences were viewed and exported using *CodonCode aligner* (version 8.0.2). Sequences were blatted against the human genome (hg38) and alignment visually inspected for confirmation of validation.

2.2.11 Development and release of *ODER*

The method for the detection of ERs was released on the Bioconductor platform as the package, *ODER*. *ODER*'s development was led by me with assistance from a member of the Ryten lab, Emmanuel Olagbaju. As part of its development, *ODER* incorporates new features to facilitate the use of the pipeline. In order to output ERs as a count matrix, the mean coverage across each ER is calculated using the *megadepth* package (59). The pipeline has also been modified to accept stranded BigWig files as input. In such cases, 2 sets of ERs will be defined (one set derived from each strand), then merged together before the ER optimisation step. In order to more accurately associate ERs to genes, *ODER* enables the filtering of genes above a user-inputted threshold expression value in a particular tissue. The gene expression data used were downloaded from GTEx v8 (https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz).

2.3 Results

2.3.1 Optimising the tissue-specific, annotation-agnostic detection of transcription

Pervasive transcription of the human genome, the presence of pre-mRNA even within polyA-selected RNA-sequencing libraries and variability in read depth complicates the identification of novel exons and transcripts using RNA-sequencing data (60). With this in mind, I used a set of exons with the most reliable boundaries (namely all exons from Ensembl v92 that did not overlap with any other exon) to calibrate the detection of transcription from 41 GTEx tissues. Of available annotation databases, Ensembl was selected as it is one of the most commonly used and comprehensive annotation providers. I used the tool, derfinder, to perform this analysis (53). However, I noted that while derfinder enables the detection of continuous blocks of transcribed bases termed expressed regions (ERs) in an annotation-agnostic manner, the mean coverage cut-off (MCC) applied to determine transcribed bases is difficult to define and variability in read depth even across an individual exon can result in false segmentation of blocks of expressed sequence. Therefore, in order to improve our analysis and define ERs more accurately, I applied derfinder, but with the inclusion of an additional parameter termed the max region gap (MRG), which merges adjacent ERs (see detailed Methods). Next, I sought to identify the optimal values for MCC and MRG using our learning set of known, non-overlapping exons.

This process involved generating 506 transcriptome definitions for each tissue using unique pairs of MRCs and MRGs, resulting in a total of 20,746 transcriptome definitions across all 41 tissues. For each of the 20,746 transcriptome definitions, all ERs that intersected non-overlapping exons were extracted and the absolute difference between the ER definition and the corresponding exon boundaries, termed the exon delta, was calculated (Figure 2.3a). I summarised the exon delta for each transcriptome using two metrics, the median exon delta and the number of ERs with exon delta equal to 0. The median exon delta represents the overall accuracy of all ER definitions, whereas, the number of ERs with exon delta equal to 0 indicates the extent to which ER definitions precisely match overlapping exon boundaries. The MCC and MRG pair that generated the transcriptome with the lowest median exon delta and highest number of ERs with exon delta equal to 0 was chosen as the most accurate transcriptome definition for each tissue. Across all tissues, 50-54% of the ERs tested had an exon delta = 0, suggesting I had defined the majority of ERs accurately. Taking the cerebellum as an example and comparing ER definitions to those which would have been generated applying the default derfinder parameters used in the existing literature (MCC: 0.5, MRG: None equivalent to 0), I noted an 96bp refinement in ER size, equating to 67% of median exon size (Figure 2.3b, 2.3c). In summary, by using known exons to calibrate the detection of transcription, I generated more accurate annotation-agnostic transcriptome definitions for 13 regions of the CNS and a further 28 human tissues.

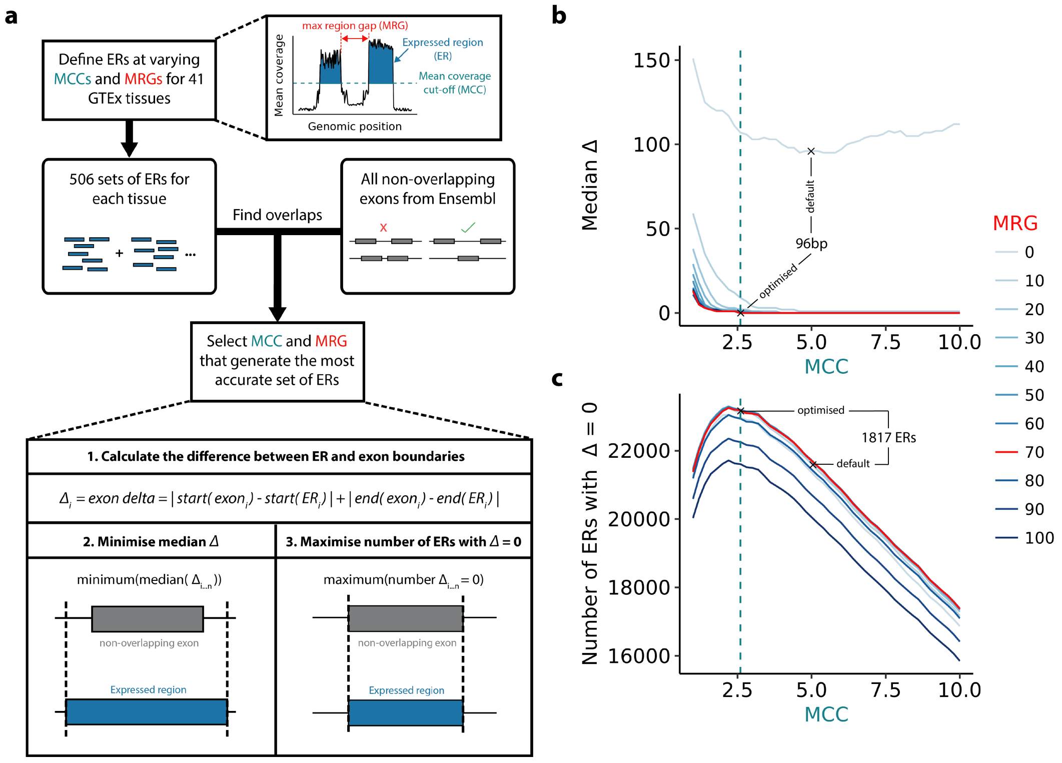


Figure 2.3 Optimisation of the detection of transcription. a) Transcription in the form of expressed regions (ERs) was detected in an annotation agnostic manner across 41 human tissues. The mean coverage cut-off (MCC) is the number of reads supporting each base above which that base would be considered transcribed and the max region gap (MRG) is the maximum number of bases between ERs below which adjacent ERs would be merged. MCC and MRG parameters were optimised for each tissue using the non-overlapping exons from Ensembl v92 reference annotation. b) Line plot illustrating the selection of the MCC and MRG that minimised the difference between ER and exon definitions (median exon delta). c) Line plot illustrating the selection of the MCC and MRG that maximised the number of ERs that precisely matched exon definitions (exon delta = 0). The cerebellum tissue is plotted for (b) and (c), which is representative of the other GTEx tissues. Green and red lines indicate the optimal MCC (2.6) and MRG (70), respectively.

2.3.2 Novel transcription is most commonly observed in the central nervous system

To assess how much of the detected transcription was novel, ERs were categorised with respect to the genomic features with which they overlapped as defined by the Ensembl v92 reference annotation (exons, introns, intergenic; Figure 2.2a). Those that solely overlapped intronic or intergenic regions were classified as novel. I discovered 8.4 to 22Mb of potentially novel transcription across all tissues, consistent with previous reports that annotation remains incomplete (18, 19). Novel ERs predominantly fell into intragenic regions suggesting that I were preferentially improving the annotation of known genes, rather than identifying new genes (Figure 2.4) Although novel transcription was found to be ubiquitous across tissues, the abundance varied greatly between tissues (Figure 2.4b & 2.4d, 2.4e). To investigate this further, I calculated the coefficient of variation for exonic, intronic and intergenic ERs. I found that the levels of novel transcription varied 3.4-7.7x more between tissues than the expression of exonic ERs (coefficient of variation of exonic ERs: 0.066Mb, intronic ERs: 0.222Mb, intergenic ERs: 0.481Mb). Furthermore, focusing on a subset of novel ERs for which I could infer the precise boundaries of the presumed novel exon (using intersecting junction reads), I found that more than half of these ERs were detected in only 1 tissue and that 86.3% were found in less than 5 tissues (Figure 2.5a). Even when restricting to ERs derived from only the 13 CNS tissues, 34.3% were specific to 1 CNS region (Figure 2.5b). This suggests that novel ERs are largely derived from tissue-specific transcription, potentially explaining why they had not already been discovered. This finding lead us to hypothesise that genes highly expressed in brain would be amongst the most likely to be re-annotated due to the difficulty of sampling human brain tissue, the cellular heterogeneity of this tissue and the particularly high prevalence of alternative splicing (46). As predicted, the quantity of novel transcription found within brain was significantly higher than non-brain tissues (p-value: 2.35e-10) (Figure 2.4e & 2.4f). In fact, ranking the tissues by descending Mb of novel transcription demonstrated that tissues of the CNS constituted 13 of the top 14 tissues. Interestingly, the importance of improving annotation in the human brain tissue was most apparent when considering purely intergenic ERs and ERs that overlapped exons and extended into intergenic regions (Figure 2.4d & 2.4e). This observation raised the question of which factors were most important in determining whether a gene was re-annotated (connected to a novel ER). I used logistic regression to find genic properties, such as measures of structural complexity and specificity of expression to brain, that significantly changed a gene's likelihood of re-annotation. I also accounted for factors which might be expected to contribute to errors in ER identification, including whether the gene overlapped with another known gene making attribution of reads more complex. I found that the annotation of longer, brain-specific genes with higher transcript complexity were more likely to have evidence for incomplete annotation (Table 2.1). Importantly, overlapping genes were not significantly more likely to be re-annotated (taking into account gene length), suggesting that novel transcription is not merely a product of noise from intersecting genes. Taken together these findings demonstrate that widespread novel transcription is found across all human tissues, the quantity of which varies extensively between tissues. CNS tissues displayed the greatest quantity of novel transcription and accordingly, genes highly expressed in the human brain are most likely to be re-annotated.

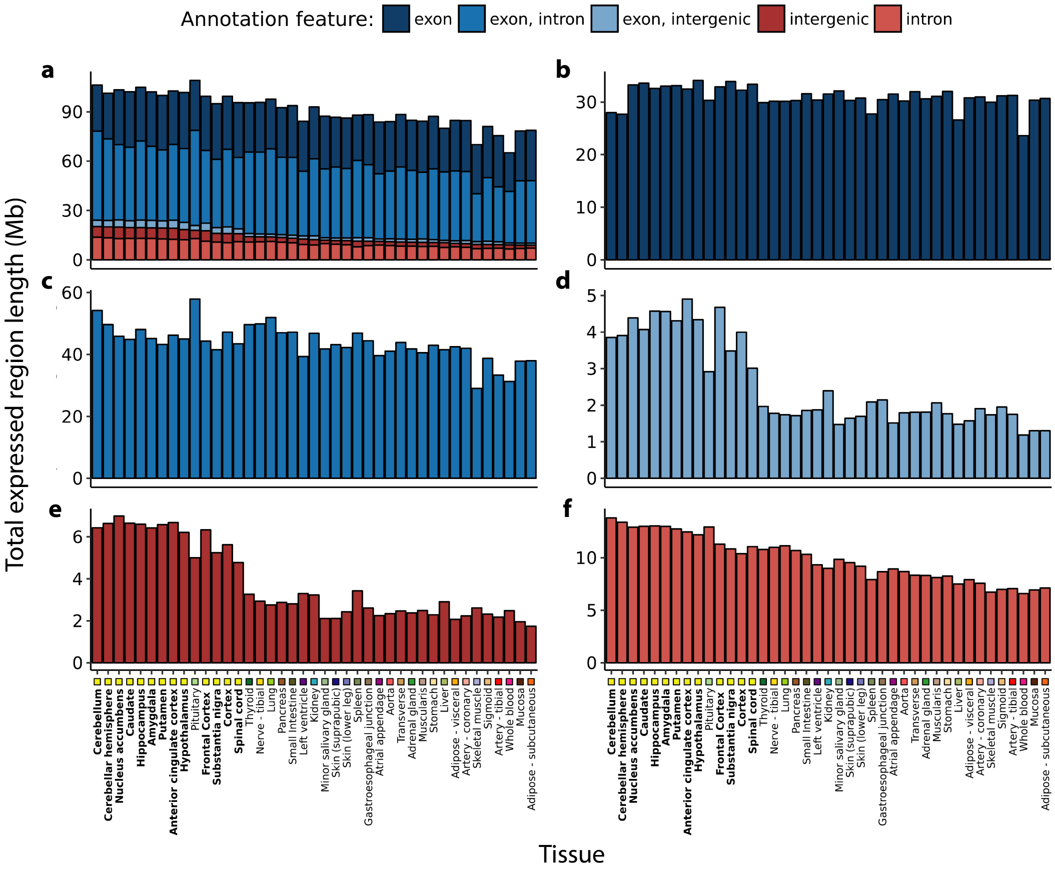


Figure 2.4 Transcription detected across 41 GTEx tissues categorised by annotation feature. Within each tissue the length of the ERs Mb overlapping a) all annotation features b) purely exons c) exons and introns d) exons and intergenic regions e) purely intergenic regions f) purely introns according to Ensembl v92 was computed. Tissues are plotted in descending order based on the respective total size of intronic and intergenic regions. Tissues are colour-coded as indicated in the x-axis, with GTEx brain regions highlighted with bold font. At least 8.4Mb of novel transcription was discovered in each tissue, with the greatest quantity found within brain tissues (mean across brain tissues: 18.6Mb, non-brain: 11.2Mb, two-sided Wilcoxon rank sum test p-value: 2.35e-10

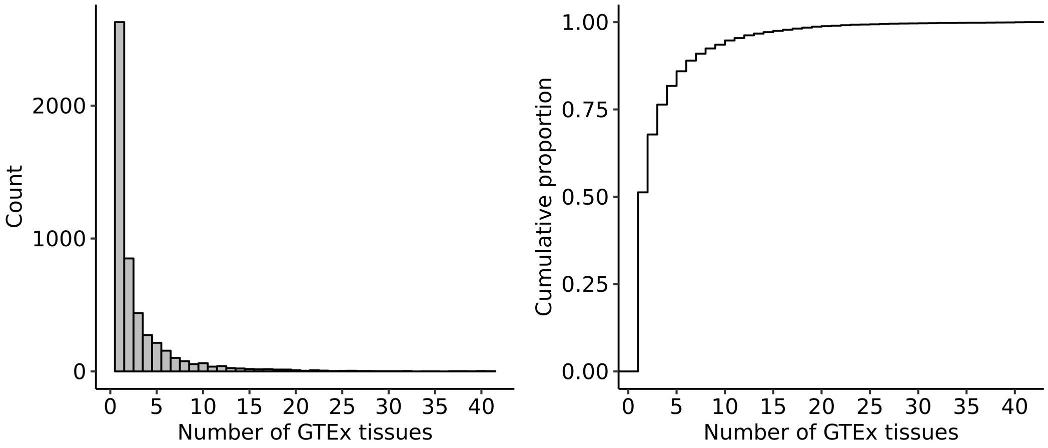


Figure 2.5 Tissue specificity of novel ERs. . Taking all intronic and intergenic ERs that were intersected by two non-overlapping split reads, we inferred the precise boundaries of this set of 5,129 unique novel ERs. We then counted the number of tissues in which these ERs were detected. The majority (51.3%) of ERs were detected in only 1 tissue and 85.9% were detected in less than 5 tissues

	Gene property	Odds ratio	p-value	Estimate
1	Brain-specific	1.10	***	0.09
2	Transcript count	1.02	***	0.02
3	Gene length	1.00	***	0.00
4	Gene biotype - protein coding	1.24	***	0.22
5	Gene biotype - lincRNA	0.96	***	-0.04
6	Gene biotype - processed pseudogene	0.86	***	-0.15
7	Gene biotype - unprocessed pseudogene	0.91	***	-0.09
8	Gene biotype - other	0.89	***	-0.11
9	Gene TPM	1.00	0.403	-0.00
10	Overlapping gene	1.00	0.773	-0.00
11	Constant	1.15	***	0.14

Table 2.1 Gene properties influencing re-annotation. Gene characteristics such as brain specificity, transcript count, gene length, mean TPM and whether the gene overlapped with another were used to assess which genes were the most likely to be identified as re-annotated. Brain-specific, longer, protein-coding genes of high transcript complexity were the most likely to be re-annotated.

2.3.3 Validation of novel transcription

A proportion of novel transcription may originate from technical variability or pre-mRNA contamination. Therefore, the reliability of detecting novel ERs across different versions of Ensembl and within an independent dataset was assessed. Firstly, I measured how many Kb of the transcription detected would have been classified as novel with respect to Ensembl v87, but was now annotated in Ensembl v92 and found that across all tissues an average of 68Kb (43-127Kb) had changed status. This value was 5.3x (3.2-10.1x) greater in every tissue compared to the Kb of ERs overlapping exons in Ensembl v87 that had become purely intronic or intergenic in Ensembl v92 (Figure 2.7a). To further assess whether this was greater than what would be expected by chance, I compared the total Kb of novel ERs entering v92 annotation for each tissue to 10,000 sets of random length-matched intronic and intergenic regions. For all tissues, the total Kb of both intronic and intergenic ERs that were now annotated in Ensembl v92 was significantly higher than the total Kb distribution of the randomised negative control regions, implying a high validation rate of novel ERs (Figure 2.6). Notably, brain regions had significantly higher Kb of ERs entering Ensembl v92 annotation from Ensembl v87 than non-brain tissues, even when subtracting the Kb of ERs leaving Ensembl v87 (p-value: 7.6e-9), suggesting the greater abundance of brain-specific novel transcription was not purely attributed to increased transcriptional noise. While our analysis of novel ERs across different Ensembl versions provided a high level of confidence in the quality of ER calling, it was limited to ERs which had already been incorporated into annotation and did not provide an overall indication of the rate of validation across all ERs. Therefore, I investigated whether our GTEx frontal cortex derived ERs could also be discovered in an independent frontal cortex dataset reported by Labadorf and colleagues (55). As expected, ERs which overlapped with annotated exons had near complete validation (>= 89%), but importantly 62% of intergenic and 70% of intronic ERs respectively were also detected in the second independent frontal cortex dataset (Figure 2.7b). While this high validation rate implied the majority of all ERs were reliably detected, I investigated whether a subset of ERs supported with evidence of RNA splicing as well as transcription would have even better rates of validation. Evidence of transcription is provided by the coverage data derived using derfinder, whilst junction reads, which are reads with a gapped alignment to the genome provide evidence of the splicing out of an intron. With this in mind, I focused on the putative spliced ERs as indicated by the presence of an overlapping junction read. Consistent with expectation, I found that ERs with junction read support had higher validation rates than ERs lacking this additional feature. This increase in validation rate for ERs with junction read support was greatest for intergenic and intronic ERs with the validation rate rising to 87% for intergenic ERs and 88% for intronic ERs (as compared to 99% for ERs overlapping exons, Figure 2.7b). Even when considering this set of highly validated ERs with junction read support, 1.7-3.8Mb of intronic and 0.5-2.2Mb of intergenic transcription was detected across all 41 tissues. Thus, in summary, the majority of novel ERs were reliably detected and validated in an independent dataset.

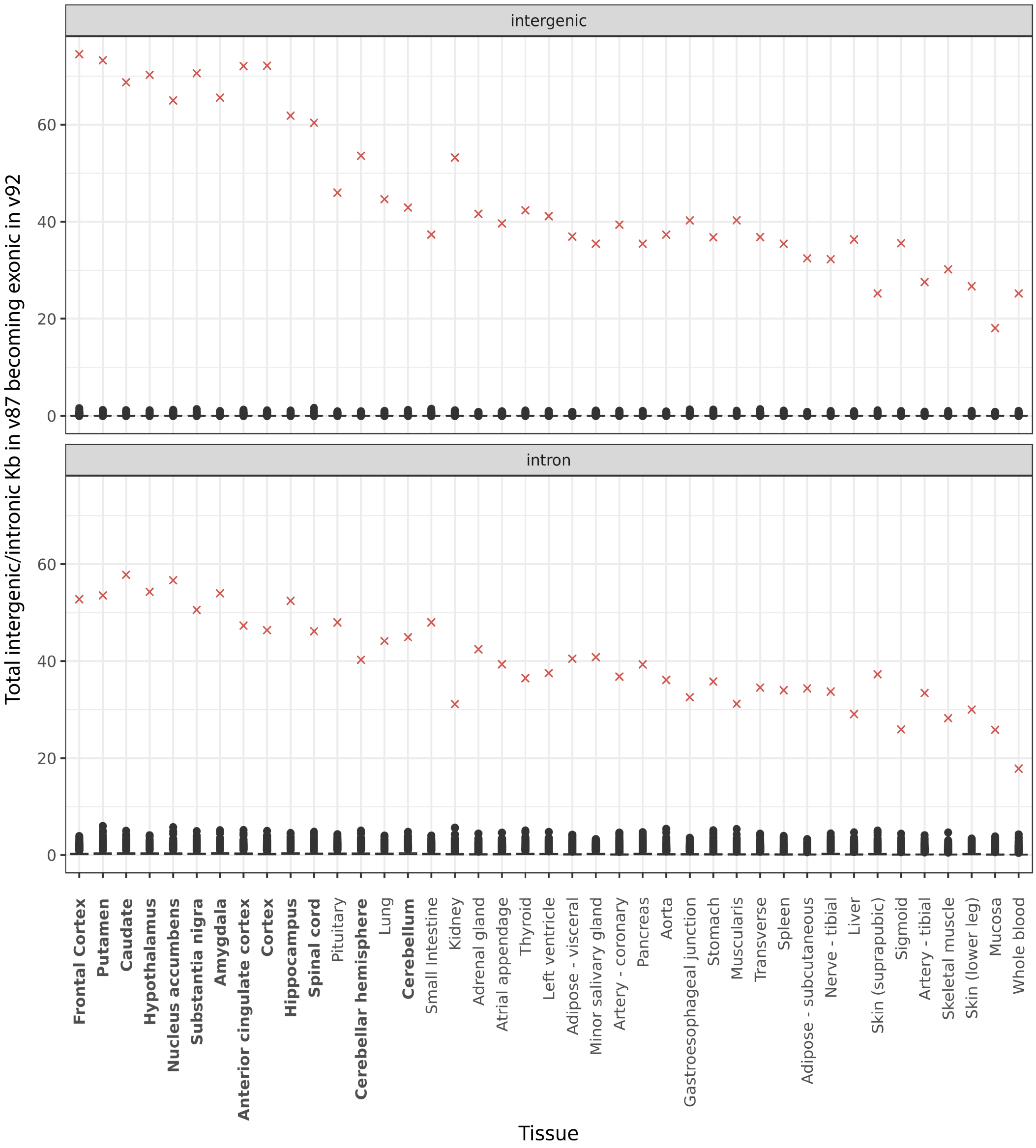


Figure 2.6 Total Kb of novel ER entering Ensembl v92 annotation compared to random, length-matched intron and intergenic regions. For each of the 41 tissues, 10,000 random sets of intron and intergenic (with respect to Ensembl v87) regions were generated and length matched to the intron and intergenic ERs derived from that tissue. For all 10,000 sets, we counted the total Kb of regions that were now exonic in Ensembl v92, shown by distributions of black dots on the graph. Red "X"'s mark the actual total Kb of novel ERs for each tissue that were validated and one-sample Wilcoxon rank sum tests were used to test whether this quantity was significantly different from the randomised sets (all p-values < 2e-16).

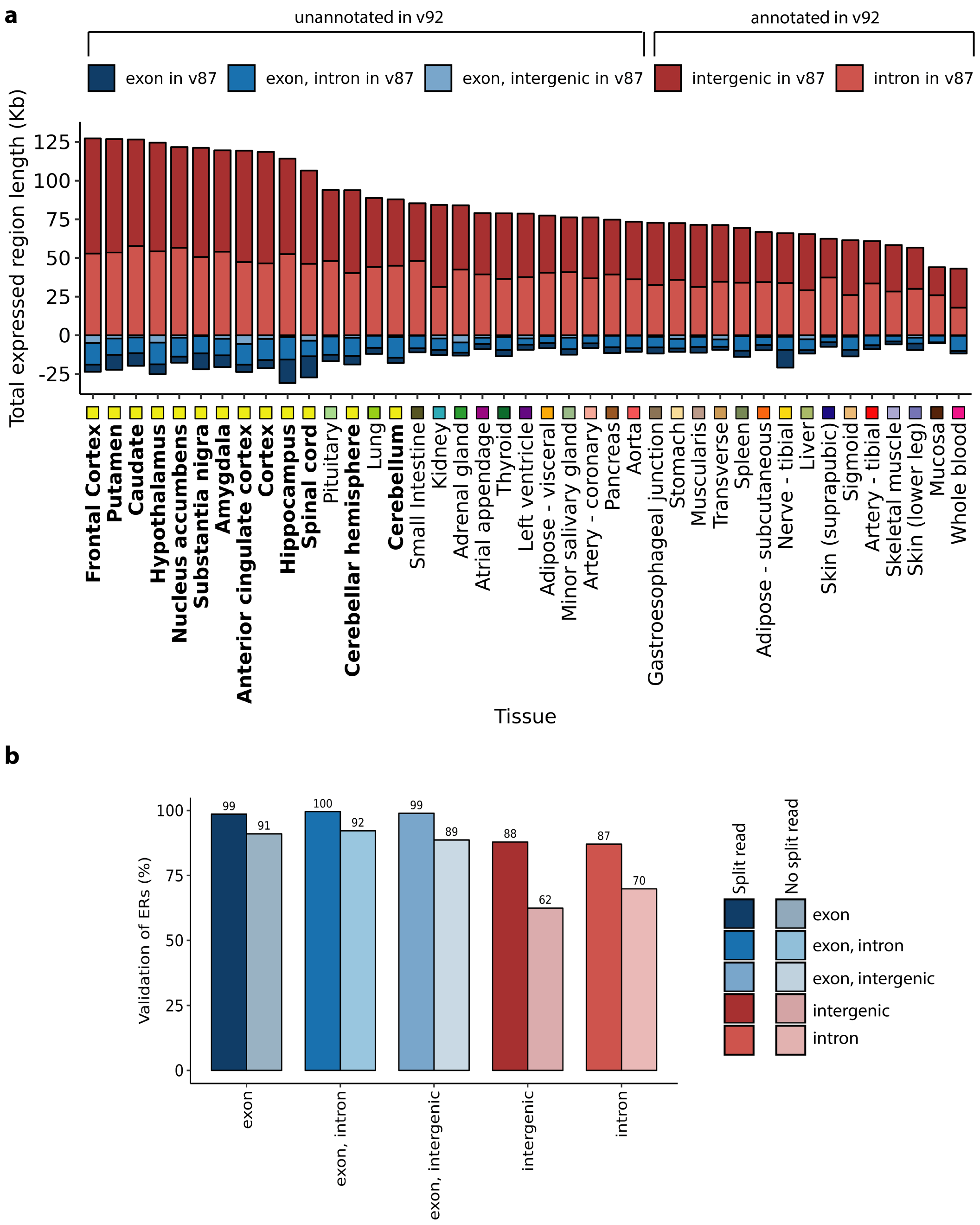


Figure 2.7 Validation of novel transcription. a) The classification of ERs based on v87 and v92 of Ensembl was compared. Across all tissues, the number of intron or intergenic ERs with respect to v87 that were known to be exonic in Ensembl v92 was greater than the number of ERs overlapping exons according to v87 that were now unannotated in v92. Tissues are plotted in descending order based on the total Mb of novel ERs with respect to Ensembl v87 that were validated (classified as exonic in the Ensembl v92). Tissues are colour-coded as indicated in the x-axis, with GTEx brain regions highlighted with bold font. b) Barplot represents the percentage of ERs seeding from the GTEx frontal cortex that validated in an independent frontal cortex RNA-seq dataset. ERs defined in the seed tissue were re-quantified using coverage from the validation dataset, after which the optimised mean coverage cut off was applied to determine validated ERs. Colours represent the different annotation features that the ERs overlapped and the shade indicates whether the ER was supported by junction read(s).

2.3.4 Novel expressed regions are likely to be functionally important within humans

Given recent reports suggesting widespread transcriptional noise and acknowledging that transcription, even when tissue-specific, does not necessarily translate to function I investigated whether novel ERs were likely to be of functional significance using measures of both conservation and genetic constraint (19, 61). The degree to which a base is evolutionarily conserved across species is dependent on its functional importance and accordingly, conservation scores have been used to aid exon identification (13). However, this measure is unable to capture genomic regions of human-specific importance. Thus, I investigated novel ERs not only in terms of conservation but also genetic constraint. Constraint scores, measured here as a context-dependent tolerance score (CDTS), represent the likelihood a base is mutated within humans (57). By comparing our detected novel ERs to 10,000 randomised sets of length-matched intronic and intergenic regions, I found that both intronic and intergenic ERs were significantly less conserved, but more constrained than expected by chance (p-value < 2e-16, Figure 2.8a). This would suggest that they have an important functional role specifically in humans. Furthermore, considering the importance of higher-order cognitive functions in differentiating humans from other species, I measured the constraint of brain-specific novel ERs separately on the basis that these ERs may be the most genetically constrained of all novel ERs identified. Indeed, I found that brain-specific novel ERs were even more constrained than other novel ERs. Another metric of functional importance is whether a region of the genome is translated into protein and notably the vast majority of all known Mendelian disease mutations fall within protein-coding regions. For this reason, I investigated whether novel ERs could potentially encode for proteins. Here, I focused on the subset of novel ERs which had evidence of splicing since the overlapping junction reads can be used to assign the precise boundaries of ERs, allowing us to confidently retrieve the DNA sequence and corresponding amino acid sequence for each novel ER. A total of 2,961 ERs covering 274Kb was found to be potentially protein coding, which represented 57% of the ERs analysed (Figure 2.8b). Amongst this set of ERs with protein coding potential, 758 ERs also fell within the top 20% of most constrained regions of the genome. These ERs connect to 694 genes, 30% of which are expressed specifically in the CNS. Overall, I discovered that novel ERs are likely to have a human-specific function. I also identified an important subset of novel ERs that have protein coding potential and are highly depleted for genetic variation in humans. Together, this suggested that at least a proportion of novel ERs are functionally significant.

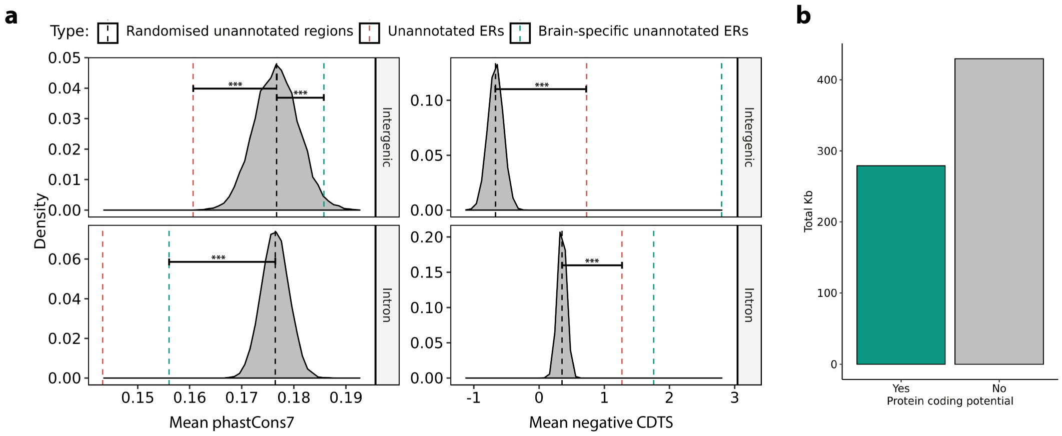


Figure 2.8 Novel ERs collectively serve an important function for humans and a proportion can form potentially protein coding transcripts. a) Comparison of conservation (phastCons7) and constraint (CDTS) of intronic and intergenic ERs to 10,000 sets of random, length-matched intronic and intergenic regions. Novel ERs marked by the red, dashed line are less conserved than expected by chance, but are more constrained. Brain–specific ERs marked by the green, dashed lines are amongst the most constrained. Data for the cerebellum shown and is representative of other GTEx tissues. b) The DNA sequence for ERs overlapping 2 junction reads was obtained and converted to amino acid sequence for all 3 possible frames. 2,168 ERs (57%) lacked a stop codon in at least 1 frame and were considered potentially protein-coding.

2.3.5 Incomplete annotation of OMIM genes may limit genetic diagnosis

Since re-annotation of genes already known to cause Mendelian disease would have a direct impact on clinical diagnostic pipelines, I specifically assessed this gene set. I found that 63% of this set of OMIM-morbid genes were re-annotated and 14% were connected to a potentially protein-coding ER, suggesting that despite many of these genes having been extensively studied, the annotation of many OMIM-morbid genes remains incomplete (Figure 2.9a). Given that OMIM-morbid genes often produce abnormalities specific to a given set of organs or systems, I investigated the relevance of novel transcription to disease by matching the human phenotype ontology (HPO) terms obtained from the disease corresponding to the OMIM-morbid gene, to the GTEx tissue from which ERs connected to that gene were derived. I discovered that 72% of re-annotated OMIM-morbid genes had an associated novel ER originating from a phenotypically relevant tissue (Figure 2.9b). This phenomenon was exemplified by the OMIM-morbid gene *ERLIN1*, which when disrupted is known to cause spastic paraplegia 62 (SPG62), an autosomal recessive form of spastic paraplegia, which has been reported in some families to cause not only lower limb spasticity, but also cerebellar abnormalities (62). I detected a cerebellar-specific novel ER that was intronic with respect to *ERLIN1*. The novel ER had the potential to code for a non-truncated protein and connected through intersecting junction reads to two flanking, protein-coding exons of *ERLIN1*, supporting the possibility of this ER being a novel protein-coding exon. Furthermore, the putative novel exon was highly conserved (phastcons7 score: 1) and was amongst the top 30% most constrained regions in the genome, suggesting it is functionally important both across mammals and within humans (Figure 2.9c). Similarly, I detected a brain-specific novel ER in the long intron of the gene *SNCA*, which encodes alpha-synuclein protein implicated in the pathogenesis of Mendelian and complex Parkinson's disease. This ER connected to two flanking protein-coding exons through junction reads (Figure 2.9d) and appeared to also have coding potential. Interestingly, while the ER sequence is not conserved within mammals (phastcons7 score: 0.09) or primates (phastcons20 score: 0.21), it is in the top 19% of most constrained regions in the genome suggesting it is of functional importance specifically in humans. I validated the existence of this ER both in silico and experimentally. The expression of this ER was confirmed in silico using an independent frontal cortex dataset reported by Labadorf and colleagues (55). Using Sanger sequencing, I validated the junctions intersecting the ER and the flanking exons in RNA samples originating from pooled human frontal cortex samples (Figure 2.10). In order to gain more information about the transcript structure in which the novel ER was contained, I also performed Sanger sequencing from the first (ENSE00000970013) and last coding exons (ENSE00000970014) of *SNCA* to the novel ER. This implied a full transcript structure containing a minimum of 609bp with the novel ER predicted to add an additional 63 amino acids (45% of existing transcript size). This example highlights the potential of incomplete annotation to both hinder genetic diagnosis since variants located in the novel ER linked to *SNCA* would not be captured using whole exome sequencing (WES).

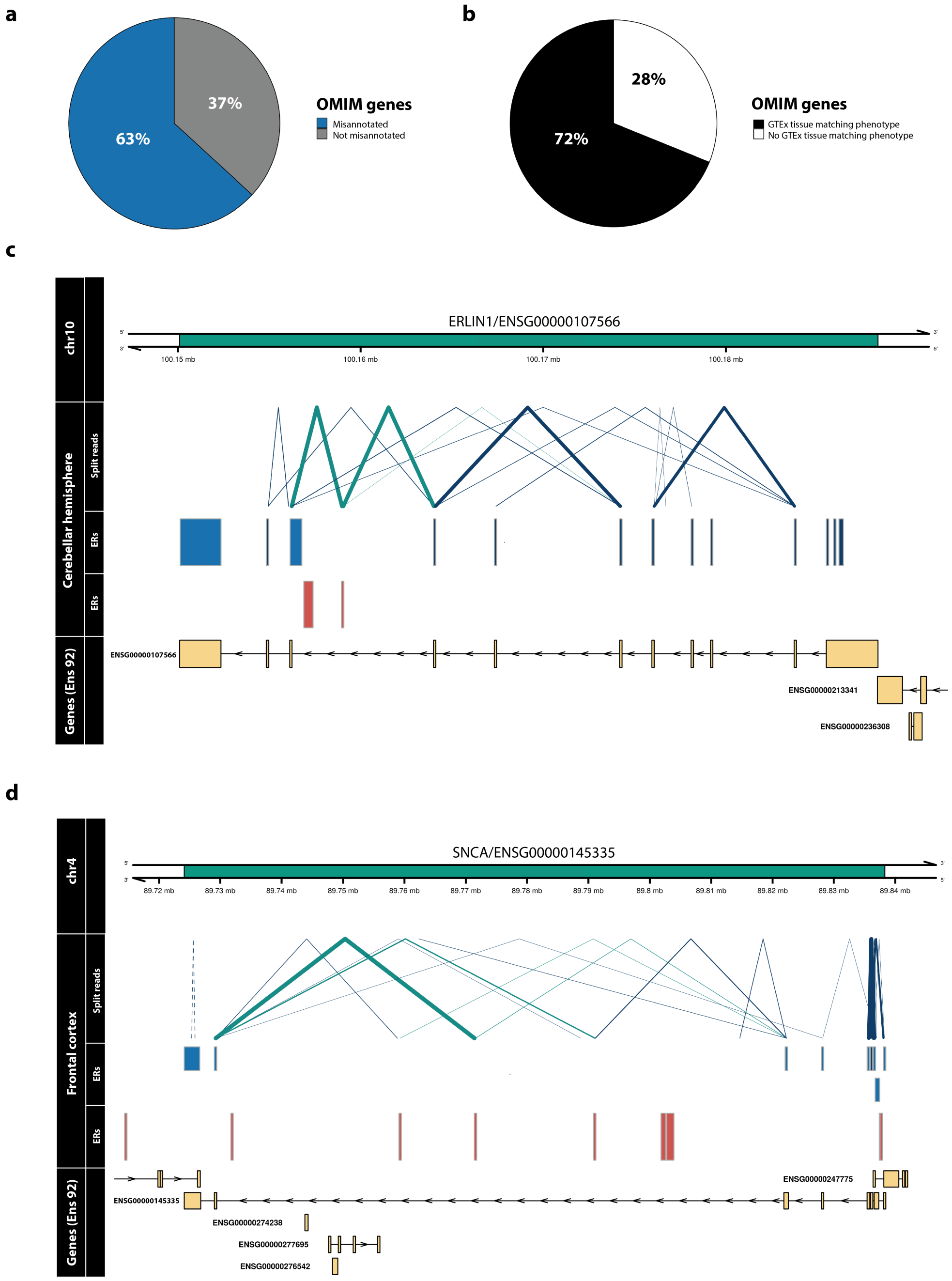


Figure 2.9 Re-annotation of OMIM genes. a) A novel ER connected through a junction read was discovered for 63% of OMIM-morbid genes. b) Comparison of the phenotype (HPO terms) associated with each re-annotated OMIM-morbid gene and the GTEx tissue from which novel ERs were derived. Through manual inspection, HPO terms were matched to disease-relevant GTEx tissues and for 72% of re-annotated OMIM genes, the associated novel ER was detected in the phenotype-relevant tissue. Visualised examples of re-annotated OMIM-morbid genes c) *ERLIN1* and d) *SNCA*. Top track represents the genomic region including the gene of interest marked in green. Second group of tracks detail the junction reads and ERs overlapping the genomic region derived from the labelled tissue. Blue ERs overlap known exonic regions and red ERs fall within intronic or intergenic regions. Blue junction reads overlap blue ERs, while green junction reads overlap both red and blue ERs, connecting novel ERs to OMIM-morbid genes. Thickness of junction reads represents the proportion of samples of that tissue in which the junction read was detected. Only partially annotated junction reads (solid lines) and unannotated junction reads (dashed lines) are plotted. The last track displays the genes within the region according to Ensembl v92, with all known exons of the gene collapsed into one "meta" transcript.

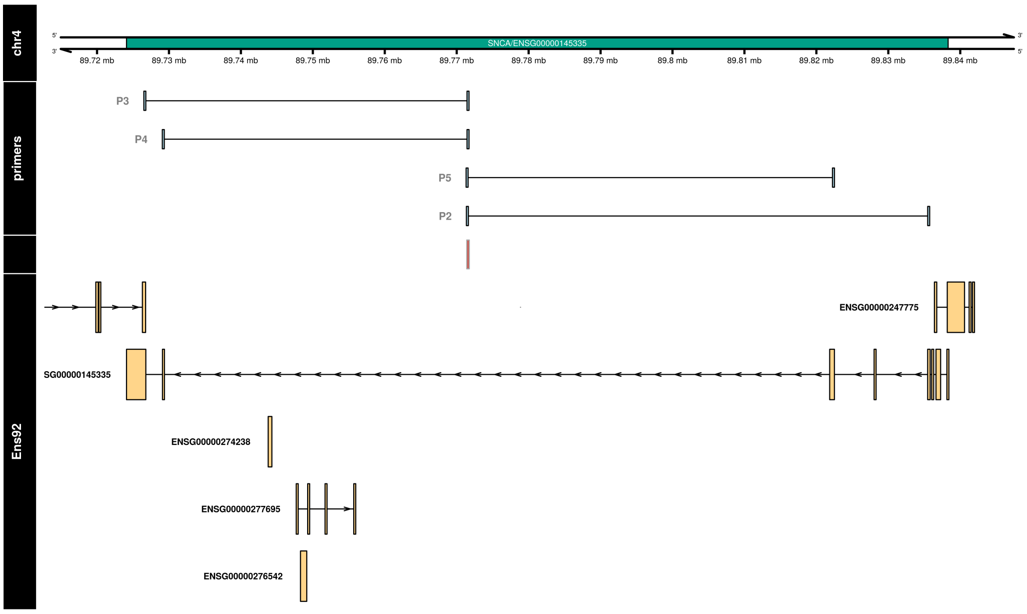


Figure 2.10 Primer locations for sanger sequence validation of *SNCA* novel exon. . The genomic locations of the of the primers used for sanger sequence validation are displayed in relation to the *SNCA* gene structure and the novel exon (in red). P4 and P5 sequenced from the novel exon to flanking exons of *SNCA*, whilst P2 and P3 sequenced from the novel exon to the first and last coding exons of *SNCA*. Full details of primer sequences are found in Table A.1

2.3.6 Automating the improvement of gene annotation using *ODER*

Although our analysis of the GTEx data set, resulted in the discovery of an abundance of novel transcription across human tissues and identified the widespread incomplete annotation of disease genes, we were aware of the limitations of this approach. More specifically, GTEx short-read RNAs-seq data represents only a small fraction of the total available RNAs-seq data available on human tissues and cells, and that in fact that through recount3 it is possible to analyse 750,000 human RNA-seq samples, with additional datasets available outside this project (63). With this in mind, I automated the approach for the detection of novel ERs (Figure 2.3a). In particular, it would be expected that that the application of this pipeline to other datasets such as those of developmental brain would yield even greater abundance of novel exons that are not yet part of annotation. Therefore, in collaboration with a colleague from the Ryten Lab, Emmanual Olagbaju, we refactored the pipeline for novel exon discovery into an R package, *ODER*, which is publicly released on Bioconductor. *ODER* takes an input BigWig and junction files from RNA-seq and outputs a set of putative novel exons. Broadly, *ODER* is comprised of 2 steps; defining and optimising the definition of ERs, then annotating ERs with respect to existing gene annotation (Figure 2.11a). It is our hope that the automation and public release of ODER will facilitate the improvement of gene annotation, through users who apply it *ODER* to their own and public RNA-seq datasets. In order to streamline the application of *ODER* for genetic analyses downstream of ERs, several new features have been added to the pipeline (Figure 2.11b). First, we have provided functionality for the conversion of the outputted novel ERs into a count matrix, calculating the average coverage across each ER as the input for each element. This is designed to be convenient for users intending to run differential expression analyses using the novel ERs. Secondly, we have enabled the association of ERs to genes, not only through junction and distance information, but also the expression of a gene in the tissue of interest. For example, if two genes lie within close proximity to a detected novel ER, *ODER* can filter for genes that are expressed in a tissue of the user's choosing, which will improve the ability to associate ERs to genes accurately. Thirdly, *ODER* permits the input of stranded BigWig files to define ERs more accurately. Finally, the output of *ODER* has been designed to be compatible with the method developed within the Ryten lab by Sid Sethi, *F3UTER* (https://github.com/sid-sethi/F3UTER). *F3UTER* uses machine learning to classify whether novel ERs that fall on the 3' end of a gene are likely to be 3'UTRs using various features derived from the RNA-seq data (64). This may be useful for users who wish to study the diversity of and variation in 3'UTRs. Overall, these improvements to *ODER* allow users' to have more versatility when conducting downstream analyses on novel ERs. Together, the public release, automation and improvements to the pipeline to discover novel ERs available through *ODER* are likely to make it a more widely used and versatile tool.

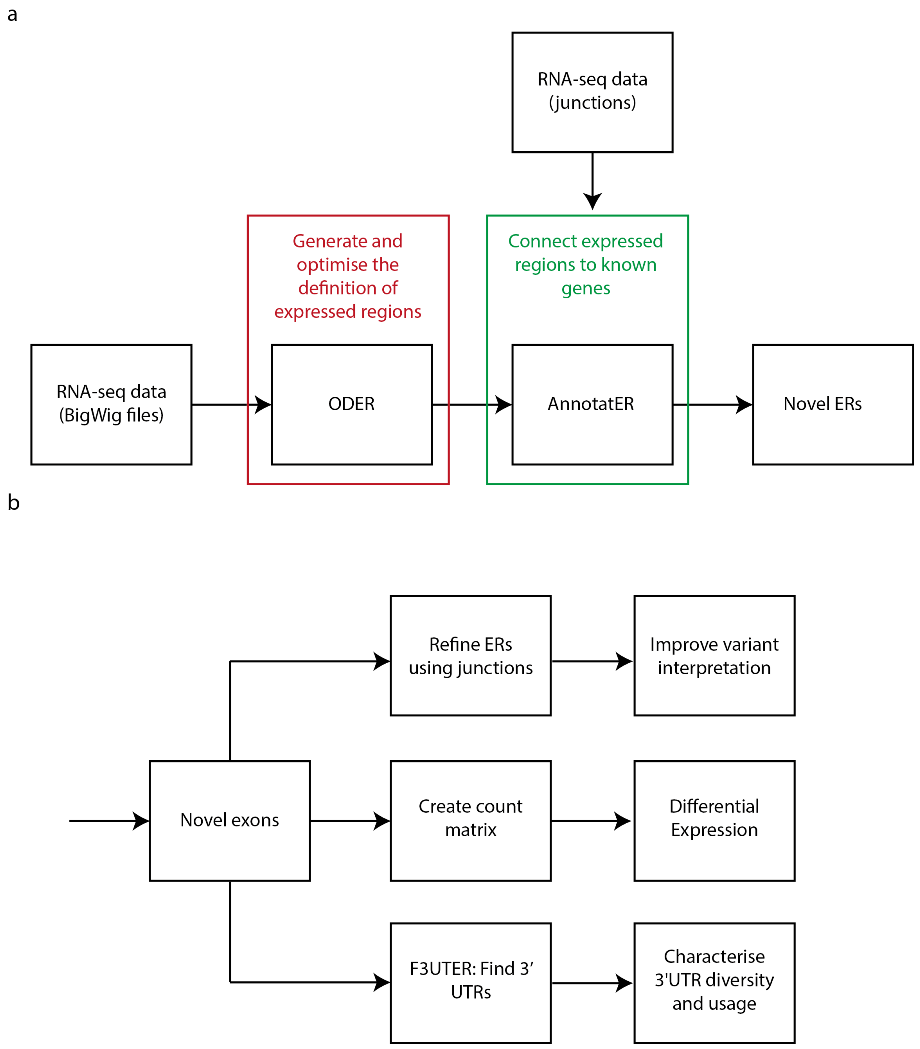

a
RNA-seq data (junctions)
Generate and optimise the definition of expressed regions
Connect expressed regions to known genes
RNA-seq data (BigWig files)
ODER
AnnotatER
Novel ERs
b
Refine ERs using junctions
Improve variant interpretation
Novel exons
Create count matrix
Differential Expression
F3UTER: Find 3′ UTRs
Characterise 3'UTR diversity and usage

Figure 2.11 Automating the improvement of gene annotation using *ODER*. a) *ODER* takes as input coverage and junctions from RNA-seq data. The first step of *ODER* is to load in the coverage data, then generate ERs across multiple MCCs and MRGs. The pair of MCCs and MRGs that minimise the exon delta are selected and associated ERs are taken forward as those with the optimal definitions. Next, ERs annotated with respect to existing annotation. Those that fall outside the boundaries of genes in intergenic regions or between exons in intronic regions are termed novel ERs. Then, 2 approaches are used to connect ERs to genes. ERs can be connected to a gene through an overlapping junction or if the ER is with a certain distance to a gene. After applying *ODER*, there are several downstream applications for the outputted novel ERs. Here, I give 3 examples of downstream applications that *ODER* has included functionality to facilitate. First, *ODER* includes functionality to further refine the ER definitions by closing ERs to the boundaries of junctions, which is particularly useful for downstream application, whereby the ERs boundary definitions are paramount, such as variant interpretation in the context of rare disease. Second, ERs definitions can be converted into a count matrix using *ODER*, the standard format for commonly used, downstream analyses such as differential expression. Third, the output of ODER has been designed to function as input into *F3UTER*, which can be leveraged to study 3’ UTR diversity.

2.4 Discussion

In recent years, the use of next-generation sequencing has changed the landscape of clinical genetics. WES and to a lesser extent WGS are becoming key components of diagnostic testing and have dramatically accelerated the discovery of new disease-causing genes. However, recent analyses predict that there is a finite pool of disease-causing genes (2). With the reducing number of potential disease genes left to discover, genetic diagnosis will become more reliant on the accuracy and completeness of the annotation of known disease-related genes. Here, I build on existing resources to develop a method to accurately detect novel transcription in an annotation-agnostic manner, then connect novel ERs to known genes and ultimately, improve the annotation of 63% of OMIM-morbid genes.

In order to improve the confidence in the discovered novel transcription, I performed 3 types of validation: i. assessing the amount of novel transcription that has become annotated in newer versions of Ensembl, ii. using an independent dataset and iii. for select genes, confirming novel exon existence using sanger sequencing. I recognise that given annotation databases are continually updated, it would be useful to replicate this first validation strategy across multiple, more recent annotations of Ensembl and across gene annotation originating from other sources such as GENCODE, RefSeq and AceView. However, through the use of multiple validation strategies, I provide evidence to suggest that at least a substantial proportion of the discovered novel transcription is likely to be real.

I find that the majority of probable novel exons detected have a restricted expression pattern, which is often disease-relevant and significantly more abundant in brain. Furthermore, since our approach does not rely on conservation across species to annotate novel exons, I am able to identify ERs which are likely to be of human-specific importance. Using constraint scores generated from aligning 7,794 human genomes and PhastCons conservation scores I find that collectively the probable novel exons, while not necessarily conserved, are depleted for genetic variation within humans suggesting that they are potential sites for pathogenic variation. The putative tissue-specific origin and human-specific functions of the novel transcription detected also provides a reasonable explanation for their omission from existing annotation databases and the abundance of novel transcription in human brain. The practical difficulty of accessing the brain reduces the number of available brain-specific datasets and its higher transcriptomic diversity is known to generate a higher number of brain-specific transcripts. In addition, I find that brain-specific ERs have the highest constraint scores, emphasising their specific importance in humans. Together, these factors suggest that the resource I have generated will have the greatest impact on the diagnosis of neurogenetic disorders.

The advent of long-read RNA-seq technologies such as those provided by Oxford Nanopore and Pacific Biosciences enable the accurate capture of full-length transcript structures (65, 66). Such technologies hold considerable promise for the future improvement of gene annotation databases and in fact, have already been applied to better capture transcriptomic complexity of large, complex genes such as *TTN* and *CACNA1C* (20, 67). However, to date, whole transcriptome long-read RNA-seq has been limited by its cost and related to that its sequencing depth and the range of public data sets on which it has been applied. In contrast, short-read RNA-seq has already been implemented at a transcriptome-wide scale on a large range of individuals, tissues and cell-types. With this in mind, it is likely that for the forseeable future, short-read RNA-seq will still be a valuable resource for improving gene annotation. For this reason, I release the method for generating the novel exons developed in this study as a Bioconductor package, *ODER*. It is my hope that this will facilitate the re-application of *ODER* on short-read RNA-seq data to further improve the annotation of disease-causing genes. I anticipate that this will have the most potential when applied to datasets that still remain less well-covered in gene annotation databases, including data generated from single-cell RNA-seq and samples across various stages of development. (68–70). In addition, I release the novel exons discovered in this study via a dedicated web resource, *vizER*, which enables individual genes to be queried for incomplete annotation as well as the download of all novel exon definitions. I anticipate this will serve as an important resource for clinical scientists in the diagnosis of Mendelian disorders.

Chapter 3

Detection of pathogenic splicing events from RNA-sequencing data using *dasper*

3.1 Introduction

Next-generation sequencing has greatly accelerated the discovery of novel gene-to-disease associations (2, 71). As a result, whole exome sequencing (WES) and more recently, whole genome sequencing (WGS) are increasingly incorporated into the genetic diagnostic routine. However, it is estimated that the success rate of such DNA-sequencing approaches in Mendelian diseases is plateauing at 35-50% (3, 5, 72). To an extent, this is due to the challenges of interpreting genetic variation beyond those that alter protein sequence or DNA structure (73, 74). In particular, non-coding regulatory variants remain difficult to assess and are more likely to be classified as variants of unknown significance (VUSs), as compared to coding variants for which more analytic approaches exist (75). Pathogenic variants that impact splicing are one class of non-coding variation, which are likely to account for a significant proportion of unsolved cases (76). The splicing machinery is tightly regulated by numerous cis and trans signals; this complexity is crucial for generating transcript and phenotypic diversity, but also increases the likelihood that genetic variation will disrupt splicing (38, 77). In fact, variants distributed in non-coding regions of the genome disproportionately affect splicing, often through disruptions to intronic splicing enhancers, silencers or recognition sequences. Furthermore, aberrant splicing has been shown to be a primary cause of rare diseases, with an estimated one third of pathogenic variants impacting splicing (40, 41).

Given the prevalence of unsolved rare disease patients with putative genetic causes through disruptions to splicing, there has been growing interest in the application of RNA-sequencing (RNA-seq) for diagnostics to directly measure transcriptome-wide splicing (22). Using RNA-seq, researchers can obtain a functional readout of splicing levels, gene expression and allele-specific expression (ASE) in patients relative to unaffected controls. This enables the discovery of aberrant molecular products, which can be used to resolve the list of candidate genes and variants identified through WGS/WES to an actionable number. Aberrant RNA-level events discovered in this way can be used to re-prioritise VUS, leading to assignment of pathogenicity. Previous publications have demonstrated the promising utility of RNA-seq for diagnostics, with success rates ranging from 7.5-21% for patients with no candidate genes after WES and/or WGS (23, 24, 27, 28). In principle, information on splicing, gene expression and ASE obtained from RNA-seq all have diagnostic potential. However, in practice the majority of genetic diagnoses made through RNA-seq have involved detection of aberrant splicing and/or aberrant expression (23–25).

Since the first systematic application of RNA-seq for diagnostics by Cummings and colleagues in 2017, there has been growing interest in developing methods to detect pathogenic RNA events in rare disease patients (33–35, 78). Although numerous tools exist to perform differential splicing analysis, almost all are designed to identify global transcriptional differences between moderate-to-large case-control cohorts (32, 79). Few are specialised for genetic diagnosis, where success relies on distinguishing a pathogenic splicing event in a single patient (N of 1). Improvements to the methodologies to detect pathogenic splicing events will relieve clinical scientists of the requirement for manual curation, permitting the wider implementation of RNA-seq-based approaches within accredited diagnostic laboratories and increasing diagnostic success.

Here, I introduce *dasper*, a method which integrates disruptions in both exon-exon junction and base pair level coverage data through machine learning to detect aberrant splicing events in patient samples. I find that *dasper* detects pathogenic splicing events with greater accuracy than existing methods. After applying an OMIM-morbid gene filter, *dasper* is able to rank true pathogenic splicing events in the top 10 most aberrant splicing events. Furthermore, *dasper* is designed with diagnostic applications in mind and includes functionality to visualize candidate genes in the form of sashimi plots for manual inspection (Figure 3.2). Finally, I demonstrate that *dasper* is able to effectively leverage publicly-available control RNA-seq datasets, enabling

RNA-seq to be a more cost-effective, standardized solution for diagnostics. *dasper* is released as an R package on Bioconductor (http://www.bioconductor.org/packages/dasper) and it is my hope that its use will improve the detection and interpretation of pathogenic splicing events and, ultimately, the diagnostic yield for rare disease patients.

3.2 Methods

3.2.1 Patient samples

RNA-sequencing was performed on a total of 55 individuals. Written informed consent was obtained for all subjects in accordance with the Declaration of Helsinki protocols and experimental protocols approved by local institutional review boards. 16 of these were genetically diagnosed Mendelian disease patients with known pathogenic splicing variants detailed in Table 3.1. The remaining 39 samples were used as in-house controls.

Pathogenic variants were classified by their proximity to annotated acceptor or donor splice sites. Those within 10bp of an acceptor or donor site were classified as "acceptor" or "donor" variants respectively, whilst those further than 10bp away were termed "deep intronic".

3.2.2 Fibroblast culture and RNA extraction

Fibroblast cell lines cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum and 0.05 g/ml uridine. Fibroblasts were harvested by first detaching cells using TrypLE Enzyme, followed by washing with Dulbecco's Phosphate Buffered Saline (DPBS) prior to storage at -80°C. Total RNA was extracted from fibroblast pellets following the manufacturer's protocol. In order to assess RNA quality, RNA integrity numbers (RIN) were measured using Agilent Technologies 2100 Bioanalyzer or Agilent 4200 Tapestation with all RIN values found to exceed 8.0.

3.2.3 RNA-sequencing, alignment and quality control of patient samples

I prepared libraries for sequencing using the Illumina TruSeq Stranded mRNA Library Prep kit by loading 50 ng of total RNA into the initial reaction; fragmentation and PCR steps were undertaken as per the manufacturer's instructions. Final library concentrations were determined using Qubit 2.0 fluorometer and pooled to a normalized input library. Pools were sequenced using the Illumina NovaSeq 6000 Sequencing system to generate 150 bp paired-end reads with an average read depth of 100 million reads per sample. Pre-alignment quality control including adapter trimming, read filtering and base correction were performed using *fastp*, an all-in-one FASTQ preprocessor (v0.20.0) (80). Reads were aligned using *STAR* 2-pass (v2.7.0) to the hg38 build of the reference human genome (hg38) using gene annotation from Ensembl v97 (81). Novel junctions discovered in the 1st pass alignment were used as input to the 2nd pass to improve the sensitivity of junction detection. Reads were required to uniquely map to only a single position in the genome. The minimum required overhang length of an annotated and unannotated junction was set to be 3 and 8 base pairs, respectively. The output BAM files underwent post-alignment QC using *RSeQC* (v2.6.4), with all samples passing quality control after manual assessment (82).

3.2.4 Control RNA-seq data

dasper analysis was conducted with two sets of controls samples; 504 GTEx (v8) fibroblast samples or a set of 55 in-house samples (including the 16 patients). GTEx v8 fibroblast junction and BigWig data was downloaded via the recount3 R package (v1.1.2) and filtered for samples without large CNVs or chromosomal duplications and deletions (SMAFRZE = “RNASEQ”) (53, 63, 83, 84). In-house RNA-seq data in the form of BAM files were converted into the BigWig format using *megadepth* (v1.08b) for input into *dasper* (v1.1.3) (59). In order to investigate the effect of changing the number of control samples used on the detection of pathogenic events, I down-sampled control numbers systematically. For GTEx control samples, 10, 20, 40, 80, 160, 320 up to a maximum of 504 samples were used. For in-house control samples, analysis was performed using 2, 4, 8, 16, 32 up to a maximum of 55 samples. For each size (N) and type of control samples, 5 iterations were executed. For each iteration, I used N randomly selected control samples of the appropriate type as input into the *dasper* pipeline. When using in-house samples, to ensure that I were not including related patients as controls, any controls with pathogenic variants matching the current patient of interest were removed prior to down-sampling.

3.2.5 Obtaining the set of OMIM-morbid genes and relevant gene panels

The full set of Online Mendelian Inheritance in Man (OMIM) morbid genes were obtained using the biomaRt R package (v2.40.5), with gene symbols taken from the Ensembl v97 database. The Genomics England panels for neuromuscular disorders (v5.9) and mitochondrial disorders (v2.12) were downloaded from the PanelApp website (https://panelapp.genomicsengland.co.uk/panels/).

Only "green" level genes with a high degree of confidence of association with disease were retained for downstream analyses.

3.2.6 LeafCutterMD

STAR outputted junctions were wrangled into a bed format for input into *LeafCutterMD* (v2.7). All 55 in-house samples were used for intron clustering. Introns were clustered matching the settings used on the LeafCutterMD documentation, namely requiring at least 50 junctions supporting a cluster and permitting introns of up to 500kb in size. Outlier intron excision analysis was performed on the 16 patient samples using default settings. Outputted p-values were standardized to ranks for comparison with the output of dasper (35).

3.2.7 *FRASER*

FRASER was run using the Snakemake pipeline *DROP* (36, 85). Patient RNA-seq data in the form of BAM files were used as input for *DROP*. Similar to *dasper*, each patient was compared to the remaining 54 patients adopting an 1-vs-all experimental design. Ensembl v97 gene annotation was used for the reference gene definitions from which junctions were annotated. *FRASER* was run with the default configuration provided by *DROP*. Importantly, this meant that only splicing events with a delta percent-spliced-in of over 0.05 and at least 10 reads in a single sample were retained for comparison. P-values were multiple test corrected using the Benjamini-Hochberg method then ranked within each sample, with a rank of 1 specifying the splicing event with the most significant p-value.

3.2.8 dasper

Figure 3.4 depicts the top-level workflow for *dasper* described in the following section. The inputs for *dasper* (v1.1.3) were junction read files (containing reads mapping with a gapped alignment to the genome) and BigWig files (which store coverage data) for control samples and the case sample of interest. Junction reads were annotated based on: i) whether their start and/or end position precisely overlapped with an annotated exon boundary, and ii) whether that junction read matched an intron definition from existing annotation as defined by Ensembl v97 (12). Using this information together with the strand, junctions were categorised as: annotated, novel acceptor, novel donor, novel combination, novel exon skip, ambiguous gene and unannotated. Annotated junctions were those that matched an existing intron definition. Novel acceptor and novel donor junctions had a single end that overlapped with a known exon boundary. Novel combination, novel exon skip and ambiguous gene junctions had both ends overlapping known exon boundaries, however the resulting introns did not match an existing intron definition as defined within Ensembl v97 (Figure 3.1). Novel combination junctions connected to exons associated with multiple transcripts, whilst novel exon skip junctions were only associated with a single transcript. Ambiguous gene junctions were connected exons originating from 2 different genes. Unannotated junctions had neither end overlapping a known exon boundary. Junctions were filtered for those that had at least 5 counts in at least 1 sample, a length between 20-1,000,000 base pairs, did not overlap any ENCODE blacklist regions and were not classified as ambiguous gene or unannotated (86). For each junction, any other junction that shared an acceptor or donor site with it was obtained to form a junction cluster. In order to normalize the junction counts to enable comparison between samples, the counts for each junction were divided by the total counts associated with its corresponding cluster.

For each junction, 3 regions of interest were defined and used to obtain coverage information, namely the intron and the two flanking exons. Exon boundaries were based on exon definitions if the end of a junction overlapped an annotated exon. Otherwise, the putative unannotated exons were presumed to be 20bp in length. Coverage across these 3 regions was loaded from BigWig files. In order to normalize the coverage for comparison between samples, the mean coverage across each of the 3 regions was divided by the total coverage across the exons of the associated gene.

I used z-scores to assess the degree to which junctions and coverage in each patient deviated from the corresponding distribution in controls. For each junction, the coverage z-score with the greatest absolute value across the 3 regions was retained, reducing the number of z-scores per junction from 4 to 2. Junctions were then split into those which had a junction count based z-score above 0 (up-regulated) and below 0 (down-regulated). An isolation forest model was fitted on the up-regulated and the down-regulated junctions separately, using the two z-scores as input. Isolation forests are an ensemble-based outlier detection method, that detect anomalies as those that require shorter paths to isolate33. The output of the isolation forest model was an outlier score per junction. Junction-level outlier scores were aggregated to a cluster-level rank in 3 steps. First, clusters that did not contain at least 1 up-regulated and 1 down-regulated junction were omitted. Then, a mean was taken of the up-regulated and down-regulated junction with the greatest outlier scores in each cluster; this formed the cluster-level outlier score. Finally, within each patient, clusters were ranked based on this cluster-level outlier score, with a rank of 1 describing the cluster that had the lowest outlier score and so was predicted to be the most aberrant.

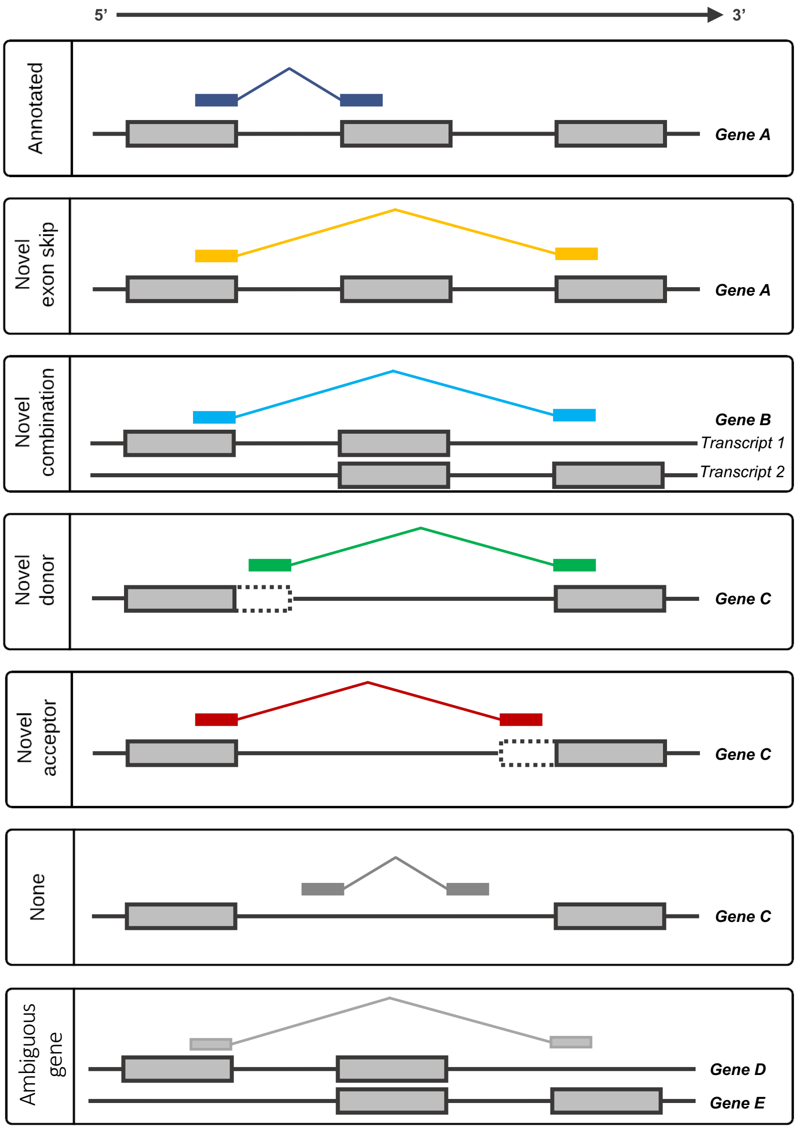


Figure 3.1 Illustration of the different categories of splicing event. Junction reads used to define Leafcutter introns were annotated based on their relationship to the annotated transcriptome (Ensembl v97). Here, the annotated transcriptome is illustrated by the grey-filled boxes. Annotated junctions have donor and acceptor splice sites that match the boundaries of an existing intron. Likewise, novel exon skip and novel combination junctions have donor and acceptor splice sites that overlap known exon boundaries derived from exons contained within the same transcript, but, they represent introns which are not found in the set of annotated introns. They are distinguished by whether or not their donor and acceptor splice sites overlap exons derived from the same transcript. Novel donors and novel acceptors are junctions where only one end (3' or 5', respectively) matches the boundary of a known exon. All novel events are considered partially annotated. Unannotated junctions ("None") have neither end overlapping a known exon. Ambiguous gene junctions are have either end overlapping different genes

3.3 Results

3.3.1 Pathogenic splicing events are characterised by abnormalities of annotated junction reads and coverage in associated regions

Previous methods to detect aberrant splicing have often focused on up-regulated novel junctions that are never or very rarely present in controls (3, 23). However, studies have demonstrated that splicing disruptions have complex consequences, which can be difficult to predict from DNA sequence data alone (43). For this reason, I first explored the consequences of pathogenic splicing variants using RNA-seq data derived from 16 Ill-characterized and deeply-sequenced patient fibroblast samples. Importantly, this cohort was selected to be heterogenous with respect to disease and variant type (Table 3.1). Patient samples were derived from individuals diagnosed with a range of neurological disease, focusing specifically on Mendelian mitochondrial disorders and rare neuromuscular conditions including Ullrich congenital muscular dystrophy (Table 3.1). All patients had diagnostically-confirmed splicing variants impacting on acceptor sites, donor sites or located deep within intronic sequence. Detailed inspection using sashimi plots of the resulting sequence data demonstrated that all pathogenic splicing events were characterized by: i) up-regulated novel junction/s (termed UJs), ii) down-regulated annotated junction/s (termed DJs), and iii) changes in coverage within the associated exonic or intronic regions. For example, analysis of RNA-seq data from an individual with a pathogenic donor splice site variant in the gene, *NDUFA4*, confirmed that this variant resulted in the generation of an UJ due to use of an novel donor site 4bp downstream of the canonical splice site (Figure 3.2) (87). However, based on the RNA-seq data I observed additional splicing changes, namely an almost complete absence of an annotated DJ, the appearance of another UJ as Ill as disruptions in coverage across the first intron (Figure 3.2). Similarly, inspection of RNA-seq data derived from an individual with a pathogenic donor splice site variant in the gene *HTRA2* (Figure 3.2), showed that as Ill as causing retention of the intron 3 with loss of the canonical splicing event (DJ), there was also a novel UJ caused by use of an novel donor site which was not previously predicted or detected35 (Figure 3.2).

	Study ID	Phenotype	Gene	RefSeq ID	Variants	Variants class	PMID
1	1.00	3-methylglutaconic aciduria, type VIII	HTRA2	NM_013247	c.906+1G>C	donor	27696117
2	2.00	Mitochondrial DNA depletion syndrome 16	POLG2	NM_007215	c.634G>A; c.689+1G>A	missense; donor	Classified by diagnostic laboratory as class 4
3	3.00	Pontocerebellar hypoplasia, type 6	RARS2	NM_020320	c.1A>G; c.613-3927C>T	missense; deep intronic	26083569
4	4.00	Mitochondrial encephalopathy and hypotonia	OXA1L	NM_005015	c.620G>T; c.500_507dup	missense; acceptor	30201738
5	5.00	Early-onset multisystem mitochondrial disease	LRPPRC	NM_133259	c.3900+1G>T	donor	26510951
6	6.00	Early-onset multisystem mitochondrial disease	LRPPRC	NM_133259	c.3147dupA; c.1582+7A>G	frameshift; donor	26510951
7	7.00	Neurological disorder	ATG7	NM_006395	c.1975C>T; c.2080-2A>G	loss-of-function; acceptor	34161705
8	8.00	Neurological disorder	ATG7	NM_006395	c.1975C>T; c.2080-2A>G	loss-of-function; acceptor	34161705
9	9.00	Cytochrome C oxidase deficiency	COXFA4	NM_002489	c.42+1G>C	donor	23746447; 29636225
10	10.00	Collagen VI-related dystrophy	COL6A1	NM_001848	c.930+189C>T	deep intronic	28424332
11	11.00	Collagen VI-related dystrophy	COL6A1	NM_001848	c.930+189C>T	deep intronic	28424332
12	12.00	Collagen VI-related dystrophy	COL6A1	NM_001849	c.930+189C>T	deep intronic	28424332
13	13.00	Collagen VI-related dystrophy	COL6A3	NM_004369	c.6210+1G>A	donor	19564581; 15563506; 25326635
14	14.00	Collagen VI-related dystrophy	COL6A3	NM_004369	c.6210+1G>A	donor	19564581; 15563506; 25326636
15	15.00	Collagen VI-related dystrophy	COL6A3	NM_004369	c.6309+3A>G	donor	20976770
16	16.00	Collagen VI-related dystrophy	COL6A1	NM_001848	c.1056+1G>A	donor	17886299; 15955946; 12840783; 10419498

Table 3.1 Details of Mendelian disease patients. The table describes the 16 patients that were used to optimise and test the performance of *dasper*. Details include the phenotype, pathogenic variants, affected gene/transcripts and the PubMed ID of the article in which each patient was diagnosed (where available).

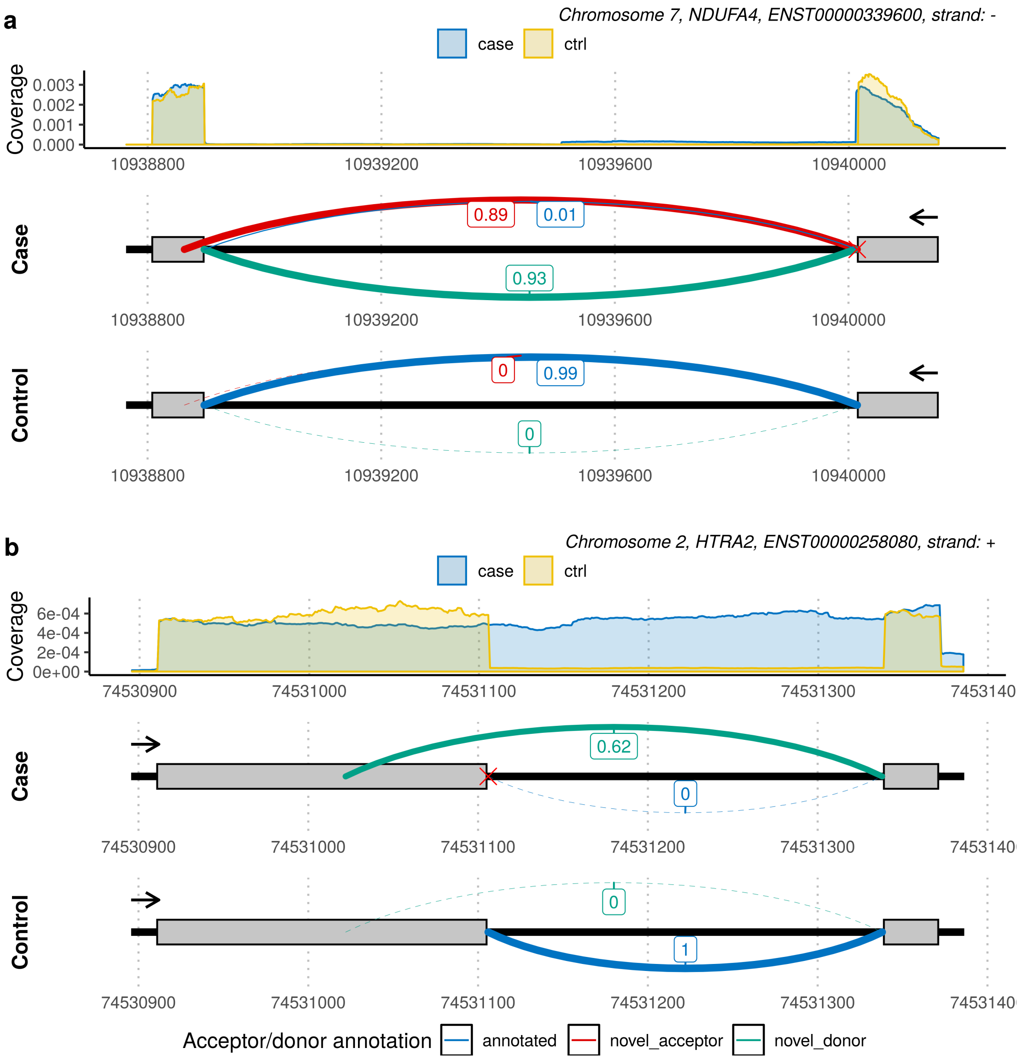


Figure 3.2 Pathogenic splicing is characterized by disruptions to junctions and coverage. a) Sashimi plots are split into two panels; the top representing the coverage and the bottom the junctions as well as gene body. Junctions are labelled with their counts and colored with respect to their annotation. The red cross represents the known pathogenic variant. The arrow represents the direction of transcription. a) In *NDUFA4*, the pathogenic splicing event can be observed through the appearance of 2 novel junctions; a novel acceptor (red) and a novel donor (green) junction, which are never found in control samples. Additionally, there is an almost complete loss of an annotated junction (blue), which is always present in control samples. Abnormalities can also be detected in the coverage data across introns associated with the aberrant junctions. First, there is a slight shift in the right-most exon boundary, which matches the donor site that is represented by the novel donor junction. Additionally, a lowly expressed, longer extension of the exon boundary is observed, which is corroborated by the annotated junction that has a normalized count of 0.01 in the patient. b) Previous studies have demonstrated that the pathogenic splicing in *HTRA2* causes an intron retention event. From the RNA-seq data, this is consistent with the loss of an annotated junction (blue) as well as a significant increase in coverage across the intron that is retained. Unexpectedly, there is also an appearance of a novel donor junction (green).

Next, I investigated the relationship between disruptions to junction usage (both UJs and DJ) and abnormalities in sequencing coverage over implied exonic and intronic regions for pathogenic splicing events. This was achieved by calculating corresponding z-scores for each of the four features of interest (UJ, UJ-related coverage, DJ and DJ-related coverage) and based on the distributions of counts and coverage in controls (50 in-house samples). I found that absolute UJ z-scores were significantly higher than DJ z-scores (median DJ: -15.21, UJ: 27.82, p-value: 0.043) and that both types of junction z-scores tended to be higher than coverage z-scores. Furthermore, I found that the correlation between junction and coverage z-scores was low (Pearson r = -0.1), suggesting that they contained distinct information. Similarly, UJ and DJ z-scores, though negatively correlated (Pearson r: -0.58), could be independently informative for detecting pathogenic splicing events 3.3. Thus, taken together this analysis suggested that pathogenic splicing events were characterized by abnormalities in UJs, DJs and nearby coverage, and that all these features could be informative.

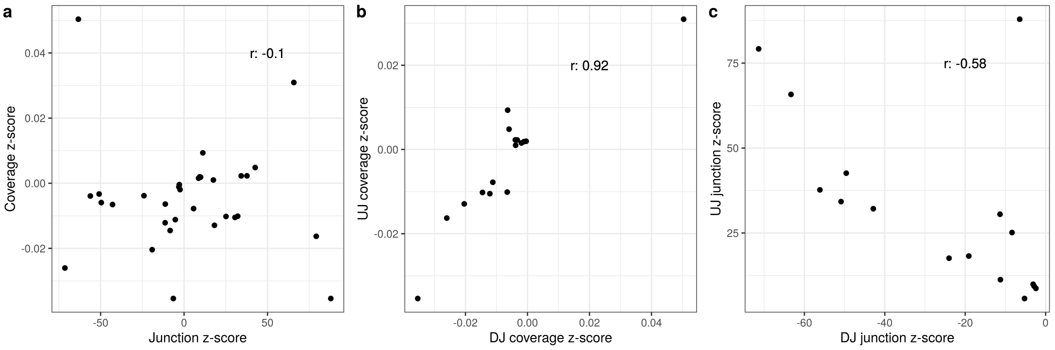


Figure 3.3 Correlation z-scores that were used as input into dasper. a) Correlation between coverage and junction z-score. **b)** Correlation between the coverage z-scores for up and down regulated junctions. **c)** Correlation between junction z-scores for up and down regulated junctions.

3.3.2 Development of a clinically accessible, machine-learning pathogenic splicing detection method

Informed by the characterization of pathogenic splicing, I next sought to improve on existing approaches for the identification of aberrant splicing through development of a new tool, *dasper*. Given that I found that pathogenic splicing variants generate both DJs and UJs within a junction cluster, *dasper* explicitly requires each splicing event to have both features, reducing the search space for pathogenic events (Figure 3.4). Furthermore, *dasper* incorporates coverage information alongside junction counts to better inform the detection of pathogenic splicing events. These key improvements are embedded within the *dasper* workflow, which begins with the input of patient RNA-seq data, and a set of user-defined RNA-seq control samples. The formats of the files required for *dasper* are standard tabular junction data and BigWigs (Methods: *dasper*). This enables easy access to large publicly available control data sets through resources such as recount2 and recount3 (63, 83). Leveraging this advantage, *dasper* includes the functionality to download GTEx control data for all clinically-accessible tissues (fibroblasts, skeletal muscle, whole blood, adipose tissue, lymphocytes), permitting the running of *dasper* with only a single patient RNA-seq analysis sourced from any of these sample types. Within *dasper* the user can then load, locally normalize and score junctions and coverage counts in patients based on their deviation from the set of controls (See methods section 3.2.8). After generating junction and coverage-related features, *dasper* applies an outlier detection method, namely an isolation forest, to aggregate junction and coverage scores in a single metric describing the aberrancy of each splicing event33. Notably, *dasper* permits easy interchange of the statistical models used to score junctions and coverage as well as the addition of other features, enabling further optimisation of the pipeline in future. Finally, the output of *dasper* is a ranked list of splicing events within each patient sample such that a rank of 1 represents the splicing event predicted to be most pathogenic (Figure 3.4). This is complemented by functions that enable visualisation of junctions and coverage of cases and controls in the form of sashimi plots to aid interpretation (Figure 3.2 and 3.4).

I assessed the utility of *dasper* and specifically the value of pairing the use of UJs and DJs, and incorporating coverage information to detect pathogenic splicing events, I compared the ranking of junctions generated on the basis of: i) UJs alone, ii) DJs alone, and *dasper* (UJs, DJs and coverage). This analysis demonstrated that *dasper* ranks pathogenic splicing events on average in the top 34 most aberrant, whilst use of only UJ or DJ information results in average similar ranks of 142 and 202 respectively (Figure 3.5). Overall, the use of information originating from both DJs and UJs, alongside the incorporation of coverage in *dasper* improves the detection of pathogenic splicing events.

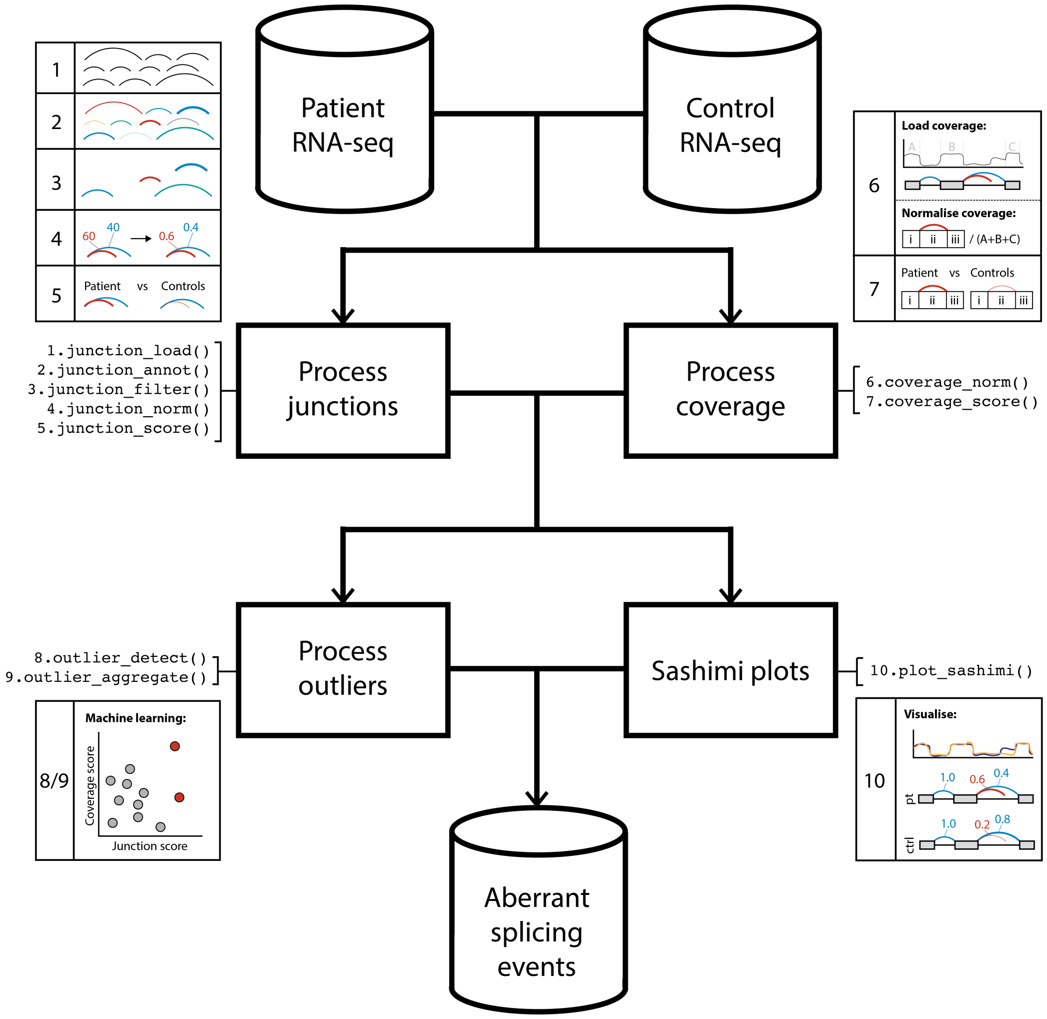


Figure 3.4 dasper applies an outlier detection method with junction and coverage information as input to detect aberrant splicing events. *dasper* takes as input RNA-seq data from a set of cases and controls. Controls can be patient samples or publicly available data, of which *dasper* includes GTEx data originating from any clinical accessible tissue. Junctions and coverage data are extracted from the RNA-seq and processed. Specifically, this involves normalizing the junction and coverage counts to permit comparison between cases and controls. Then, scoring junctions and coverage by the deviation of their counts from the corresponding count distribution in control samples. These scores are aggregated using an outlier detection model. For each patient, the outputted outlier scores are ranked, generating a list of all splicing events in each patient ranked by their aberrancy. A rank of 1 specifies the most aberrant splicing event in each patient. *dasper* includes functions to plot sashimi plots to permit manual inspection of candidate splicing events.

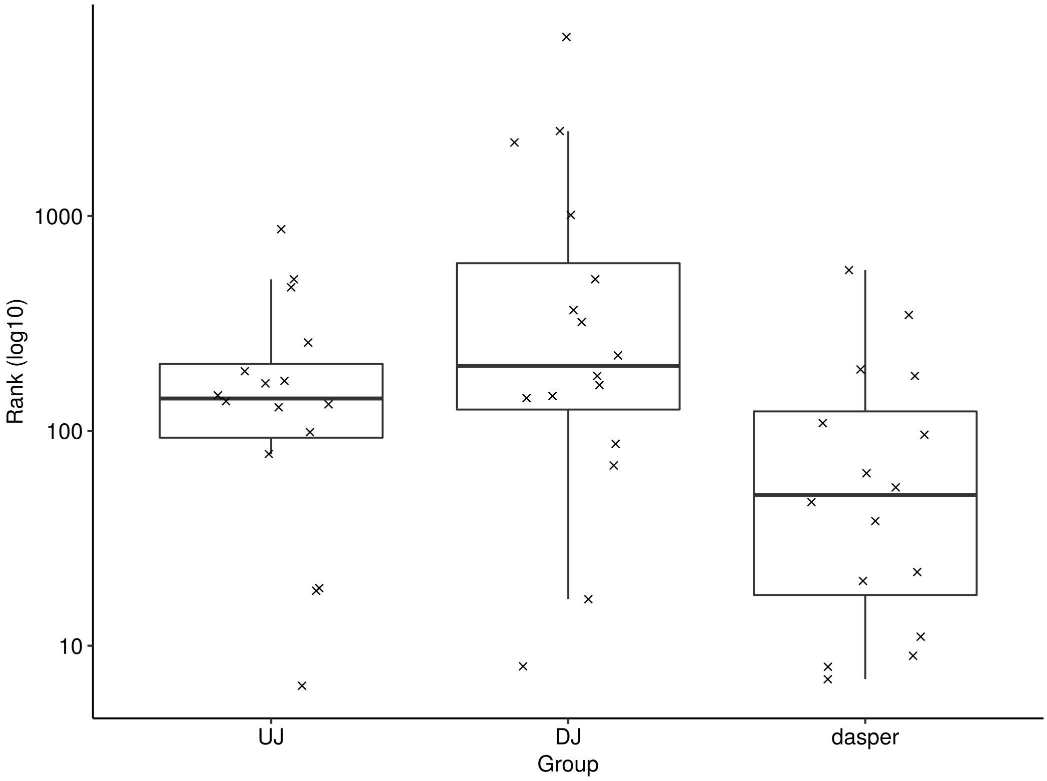


Figure 3.5 dasper applies an outlier detection method with junction and coverage information as input to detect aberrant splicing events. Boxplots displaying the rank of 16 pathogenic splicing events across varying inputs. Each point represents a pathogenic splicing event from one of the 16 patients analyzed. The x-axis shows what information has been used for the ranking, either only up-regulated junctions (UJ), down-regulated junctions (DJ) or the *dasper* method (UJs, DJs and coverage). The y-axis displays the rank outputted from *dasper*, with lower ranks specifying splicing events that are predicted to be more aberrant.

3.3.3 Comparison of *dasper* to other methods used to detect pathogenic splicing

Next, I evaluated *dasper*'s performance in comparison to existing, commonly used approaches for pathogenic splicing detection, namely the *LeafCutterMD*, *FRASER* and z-score methods (35, 36). In order to enable comparison between tools, I converted *LeafCutterMD* and *FRASER* p-values as well as z-scores to a ranking such that the lowest p-value or highest absolute z-score was assigned a rank of 1. Based on the analysis of patient-derived fibroblast samples. I found that the rankings for pathogenic splicing events produced by *dasper* were significantly lower than those generated by *LeafCutterMD* and z-score (*LeafCutterMD* wilcoxon p-value: 0.013; vs z-score: 0.0003) (Figure 3.6). However, in comparison to *FRASER*, a new method released during the writing of this thesis, *dasper* ranked the majority of pathogenic splicing events higher with the exception of 2 patients (*FRASER* wilcoxon p-value: 0.001).

Given that pathogenic splicing can vary in its difficulty of detection dependent on the type of event, I also investigated the performance of *dasper* across the different causative variant types. I found that while *dasper* detected variants at donor versus acceptor sites with similar accuracy (wilcoxon p-value: 0.482), pathogenic events caused by deep intronic variants received significantly higher ranks, indicating that they were more difficult to detect (wilcoxon p-value: 0.013) (Figure 3.6).

I recognized that the utility of *dasper* in diagnostic settings depends not only on how it compares to existing tools but on its performance in clinically-relevant contexts. To investigate this, I measured the absolute ranking of pathogenic splicing events using *dasper*. I found that pathogenic splicing events were ranked on average in the top 40 (median: 33.750) most aberrant events in each patient (Figure 3.6), but note that these ranks were obtained without any gene, variant or phenotypic level filters. Given that in diagnostic settings only genetic variants in known disease-associated genes would be considered, I re-calculated rankings after filtering for splicing events that were connected to genes within the OMIM-morbid gene set or the appropriate Genomics England panels (see detailed methods). After filtering for OMIM-morbid genes, I found that *dasper* was able to rank pathogenic splicing events within the top 10 most aberrant in each patient (median: 6.750) (Figure 3.6). The more stringent gene panel-based filtering, which not only assumes the gene has to be known to cause disease but is also linked to the patient phenotype, further reduced rankings such that pathogenic events were within the top 5 most aberrant on average (median: 2.5) (Figure 3.6). In summary, *dasper* is able to rank pathogenic splicing events such that they would be identifiable with only minimal manual curation.

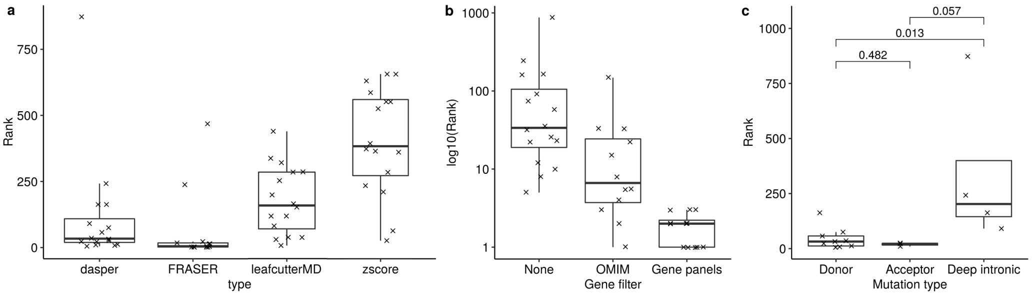


Figure 3.6 Comparison of the performace of dasper to existing pathogenic splicing detection tools. **a)** Comparison of different methods used to detect aberrant splicing. dasper ranks pathogenic splicing lower or more aberrant than existing LeafCutterMD or z-score approaches. However, dasper ranks pathogenic splicing higher than the FRASER method. The y-axis represents the rank of pathogenic splicing events, whilst the x-axis specifies the method used. **b)** Ranking pathogenic events across different gene filters. The x-axis details the sets of gene sets that have been used for filtering; either no filter, splicing events connected to OMIM-morbid genes or splicing events associated with gene panels. After applying the OMIM-morbid or Genomics England gene panel filter, pathogenic splicing events are ranked on average in the top 10 and top 5 most aberrant splicing events respectively. **c)** Pathogenic splicing events resulting from deep intronic variants are ranked higher than acceptor or donor variants, suggesting that they are more difficult to detect.

3.3.4 dasper is able to leverage publicly available control data effectively

While there is increasing evidence to show that paired patient-derived transcriptomic data can increase the diagnostic yield of WES/WGS, there remain significant barriers to implementing this approach in clinical settings. One such hurdle is the generation or identification of suitable control data. In the previous analyses, I have used 50 in-house sequenced RNA samples as controls. I are aware that sequencing this number of RNA-seq samples would incur a substantial resource burden on diagnostic labs, which may not be feasible in practice. To address this issue, I assessed the performance of dasper when using publicly available GTEx v8 data originating from 504 fibroblasts, matching the tissue of origin of patient-derived RNA-seq data in this study (51). I found that, on average, using in-house samples resulted in more accurate calling of pathogenic splicing events, when compared to the use of GTEx samples as controls. The improvement in ranking of pathogenic events when using in-house controls was observed in 14/16 patients analysed. This pattern of improvement remained true following filtering for pathogenic splicing events within known disease genes (median no filter GTEx: 90, no filter in-house: 34) (Figure 3.7. However, this analysis also demonstrated that the absolute ranking when using publicly available controls may be sufficient to be useful when applied in a more clinically-relevant manner. After limiting splicing events to only those connected to genes already implicated in genetic disease as defined in OMIM, and using GTEx controls, *dasper* was still able to rank true pathogenic splicing events in the top 25 most aberrant events (median: 24.5). Overall, this suggests that while technical variability between patient and controls samples reduces the ability to detect pathogenic splicing events, publicly available control data is a viable alternative to costly, time-consuming in-house data generation.

Next, I explored the relationship between control sample number and the power to detect pathogenic splicing events, a significant concern for implementation in a diagnostic setting whether in-house or external control data is being used. To investigate this, I applied *dasper* while randomly down-sampling the number of control samples used, analysing GTEx and in-house control data separately. As would be expected, I found that an increase in the number of controls considerably improves the detection of pathogenic splicing events using either GTEx or in-house control data (Figure 3.7). Notably, while the rate of improvement in pathogenic splicing detection greatly diminishes with increasing control number suggesting a diminishing return, it does not appear to plateau at the maximum number of available samples for either control type. This analysis would suggest that further increases in the quantity of publicly available control samples could compensate against the technical differences between patient and control sample sets. Notably, I found that the ranking when using 504 GTEx controls matched the performance of using between 8 and 16 in-house samples (Figure 3.7). In summary, it is likely an increase in sample number would improve the detection of pathogenic events for both control types.

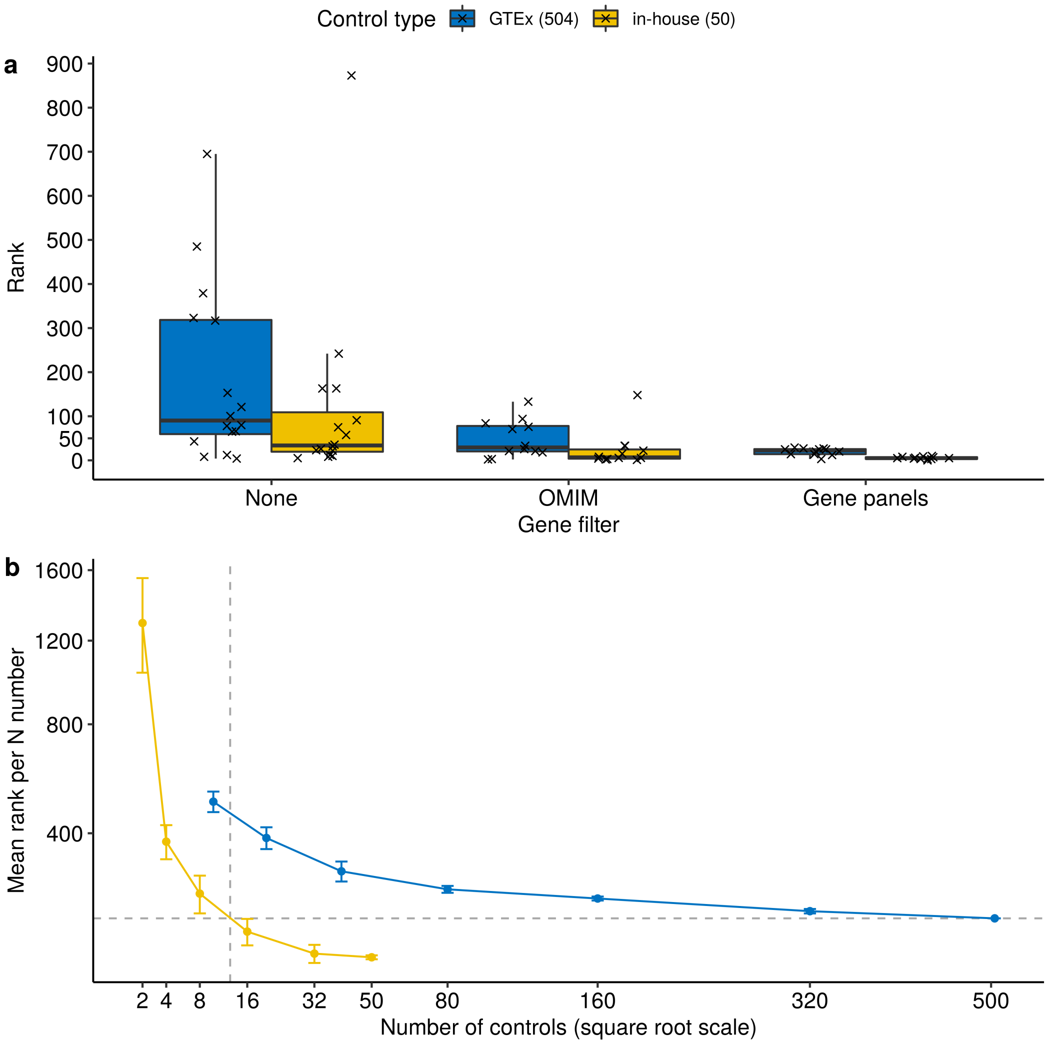


Figure 3.7 dasper is able to leverage publicly available and in-house controls effectively. a)) The rank of pathogenic splicing events across varying gene filters and control types. The colour of boxplot represents the control type used, either 504 GTEx v8 samples (blue) or 50 in-house sequenced samples (yellow). In general, in-house samples are able to detect pathogenic splicing events easier than GTEx samples. However, after applying a gene panel filter, pathogenic splicing events are detected in the top 10 splicing events for either control type. **b)** Comparison of the performance of GTEx and in-house control data for detecting pathogenic splicing events. The x-axis describes the number of controls used. The colour of the points and lines describes which control type is used, namely up to 504 GTEx fibroblasts or up to 50 in-house samples. At each N of controls analysed, the mean and standard deviation of the rank of the 16 pathogenic events for the 5 sets of randomly down-sampled controls is plotted.

3.4 Discussion

In this study, I present *dasper*, an R package released on Bioconductor that can be used to detect aberrant splicing events from RNA-seq data. Here, I use a cohort of 16 patients with known pathogenic splicing variants to inform the development of *dasper* and demonstrate its utility. Uniquely, *dasper* pairs information from DJs with UJs as well as incorporating coverage changes across a gene to improve the detection of pathogenic splicing events. *dasper* was able to rank pathogenic splicing events in the top 10 most aberrant after OMIM-morbid gene filtering. Designed with clinical accessibility and interpretation in mind, *dasper* uses standard RNA-seq data formats as input granting users flexibility to incorporate publicly available datasets as controls. Moreover, I demonstrated that *dasper* was able to leverage publicly available GTEx data effectively. Finally, *dasper* includes sashimi plot functionality to aid the manual inspection of candidate splicing events (53, 83, 88).

In comparison to existing aberrant splicing detection tools, *dasper* outperformed the *LeafCutterMD* and z-score based approaches. However, the recently released method *FRASER* ranked the majority of pathogenic splicing events more accurately than *dasper*. *FRASER* was released within the time frame of the completion of this thesis. While our benchmarking suggests that the autoencoder approach within *FRASER* does outperform *dasper*, it would be useful to confirm whether this remains the case on a larger set of patients. Furthermore, due to inherent potential of over correction when employing an autoencoder, in future studies it would be valuable to assess the performance of *FRASER* compared to *dasper* when external, publicly available controls are used.

To the best of my knowledge, this is the first study to explore the impact of splicing variant subtypes and control sample selection on the detection of pathogenic splicing events. My analyses highlighted that the selection of control sample type and number greatly impacts the power of pathogenic splicing detection. In particular, I compare the usage of two types of control data; either the use of publicly available GTEx RNA-seq data or in-house sequencing data. While I find that the use of in-house samples improves the performance of *dasper*, presumably because of a reduction in the technical differences between patient and control data, this approach is associated with increased costs and reduced flexibility, creating barriers to the use of RNA-seq pipelines in diagnostic laboratories. In contrast, the use of publicly available datasets has minimal associated costs and is highly flexible; it enables any kind of (clinically-accessible) tissue sample to be analysed for a given patient, reduces the need batch patient samples together, which would reduce turnaround times for laboratory results. Although *dasper* performed better when using in-house samples, GTEx samples still enabled pathogenic splicing events to be detected, on average, in the top 25 most aberrant after applying an OMIM-morbid filter alone. This ranking was equivalent to using 8-16 in-house samples suggesting that use of publicly available data could be a viable, cost-effective alternative for the detection of pathogenic splicing. In this context, it is worth noting that public RNA-seq datasets are progressively increasing in size. In fact, for tissues such as blood, public datasets collectively could provide RNA-seq profiles for >30,000 unrelated individuals which could be meta-analysed as elegantly demonstrated by the eQTLgen consortium (88). Additionally, over 70,000 and 300,000 human RNA-seq samples are publicly released through the recount2 and recount3 projects respectively (53, 63, 83). It would helpful to assess the usage of publicly available control datasets of this size on pathogenic splicing detection, as one might expect that the increased N number within such datasets could have the potential to match the efficacy of in-house controls. Furthermore, if put into practice, the use of publicly available controls would permit different centers to use identical computational protocols for diagnoses thus enable the standardization of pathogenic splicing identification across laboratories.

Together, through leveraging in-house or publicly available datasets effectively, it is my hope that *dasper* will make RNA-seq a more affordable, effective and standardized tool for diagnostics and ultimately, lead to an increased rate of genetic diagnoses for Mendelian disease patients.

Chapter 4

Improving the diagnostic rate of patients with suspected mitochondrial disorders using RNA-sequencing

4.1 Introduction

Although individually rare, Mendelian diseases collectively affect an estimated 3.5-6% of the human population, with an estimated 80% of disorders expected to have a genetic origin (1). Establishing a genetic diagnosis in rare disease patients enables a more accurate prognosis, informs genetic counselling, can improve the management of disease symptoms and, in some cases, enables disease-modifying therapies to be administered (2). The advent of next-generation sequencing technologies has revolutionised the landscape of clinical genetics, as evidenced by the incorporation of whole-exome sequencing (WES) and whole-genome sequencing (WGS) into the diagnostic routine. These technologies have reduced the cost of sequencing a human genome and as a consequence, accelerated the number of gene-disease associations identified in recent years (71). However, even after the application of WGS, a recent report found that a genetic diagnosis was achieved in only 33% of probands across a large range of clinical phenotypes and diseases (5). While it is worth noting that diagnostic yield varied widely by disorder type (range: 0-55%), whether the disorder was thought to be entirely monogenic in origin or likely to have a complex cause (range: 11-35%) and the family structure (5), across all disease areas there remain many undiagnosed patients. This is largely because although WGS is capable of capturing the vast majority of variation within a human genome, accurate interpretation of this variation remains a major challenge for diagnostics.

Variant interpretation relies on the cumulative evidence from population-wide variant frequency data, bioinformatic predictions, functional assays, and segregation patterns (25). Functional data requires determining the consequence of a variant on RNA and/or protein abundance, structure and function (3). Given the historic focus on protein-coding regions, prediction of the consequence of non-coding variants on expression, splicing or RNA stability remains particularly challenging (37). Of all pathogenic variants, it is suggested that 30% fall in non-coding regions (40). A further 10% of exonic pathogenic variants are thought to impact on splicing and also remain difficult to interpret correctly (89). Despite the development of several computational tools to predict the effect of variants on transcription, their accuracy remains too low for diagnostic applications (23, 43, 90). Thus, often these non-coding variants and splice-disrupting variants remain as variants of unknown significance (VUSs) and require downstream functional assays for validation of their consequence (24).

RNA-sequencing (RNA-seq) has become the gold-standard method for the systematic detection of aberrant gene expression, splicing and allele-specific expression. These aberrant events can provide evidence of the functional impact of VUSs, thereby allowing the re-assignment of their pathogenicity (4). A recent study systematically assessed the efficacy of WES, WGS and RNA-seq for genetic diagnoses and found that almost 20% of the pathogenic variants they discovered required RNA-seq data to determine their causality (4). In practice, RNA-seq has been applied to patients clinically diagnosed with a variety of disorders and has been shown to improve the success rate of genetic diagnosis by 5-35%, with the vast majority of diagnoses have been made through the detection of aberrantly expressed genes or aberrant splicing events (23–25, 27, 28).

Given RNA-seq data provides a transcriptome-wide functional readout of DNA, its application can identify and characterise the structure of aberrant transcripts that drive pathogenicity at a specific locus. This can be important for the development of personalised therapies, such as the design of splice-modulating treatments, an approach which has recently been applied successfully to treat a patient with Batten's disease (44). In addition, RNA-seq can be used to identify other genetic loci outside of the pathogenic gene that modify to the disease phenotype, a situation has been estimated to occur in 5% of Mendelian disease patients (91). If discovered, such genetic modifiers have the potential to improve prognostic accuracy for patients (92). Finally, RNA-seq can be used to obtain a molecular signature of the disease process, which could be used to understand the disrupted pathways as well as discover novel disease-associated genes (93, 94).

Here, I apply RNA-seq to 60 patient samples, of which 32 are derived from undiagnosed individuals, 26 are derived from individuals with a known genetic diagnosis and 2 samples are derived from unaffected individuals. I attempted to use RNA-seq to diagnose the 32 individuals with a suspected mitochondrial disease for whom WES or WGS had been inconclusive. Mitochondrial disease represents a disorder type for which a transcriptome-wide approach would be expected to provide diagnostic utility, due to its clinical heterogeneity and the large number of known causative genes (25). I used RNA-seq to detect aberrantly expressed genes and aberrant splicing events, then integrated this information with the phenotype of the patients to discover candidate genes. In total, this resulted in the successful diagnosis of 1 patient and provided candidate genes for a remaining third of the patients. By analysing these candidate genes, I also demonstrate the utility of RNA-seq for characterising splicing events. Next, by analysing 5 of the 26 patients with known genetic diagnoses with pathogenic mutations within the gene *ATG7*, I demonstrate the ability of RNA-seq to improve our understanding of the disease mechanism within these patients. RNA-seq detected a global down-regulation of mitotic, cell-cycle and golgi pathways. Furthermore, RNA-seq was able to detect the aberrant down-regulation of the gene *VPS41*, which upon further investigation revealed a heterozygous mutation within this gene that segregated in 2 of the 5 patients. Together, I demonstrate the ability of RNA-seq to improve diagnostic rates as well as improve our understanding of the pathogenic processes in Mendelian disorders.

4.2 Methods

4.2.1 Patient and control samples

RNA-sequencing was performed on fibroblasts samples from a total of 60 individuals. Written informed consent was obtained for all subjects in accordance with the Declaration of Helsinki protocols and experimental protocols approved by local institutional review boards. This includes: i) 32 patients clinically diagnosed with a suspected mitochondrial disorder (Tables A.3, A.4, A.5, A.6) 26 patients with a known genetic diagnosis and iii) 2 unaffected individuals. Of the 26 genetically diagnosed patients, 14 patients had a Mendelian mitochondrial disorder, 7 had congenital muscular dystrophy and 5 had a ATG7-associated neurometabolic disorder. Fibroblasts from the 7 congenital muscular dystrophy patients were obtained through collaboration with Dr. Haiyan Zhou and Prof. Francesco Muntoni. Fibroblasts from the remaining 53 patients have been obtained from the collaborators within the Lilly consortium including Dr. Charu.Deshpande, Dr. Ines Barbosa, Prof. Joanna Poulton, Prof. Michael Simpson, Prof. Robert McFarland, Dr. Robert Pitceathly and Prof. Robert Taylor.

The 32 suspected mitochondrial disorder patients remained genetically unsolved after WES sequencing or panel-based analysis. The phenotype of these patients is detailed in table A.4. These patients were analysed with the aim of improving their genetic diagnosis using RNA-sequencing. All downstream analyses using *dasper* or *OUTRIDER* conducted for this reason was performed in a 1-vs-all manner, whereby each of 32 patients was independently compared to a the control group. Here, the control group used for each patient consisted of a total of 59 individuals. This comprised of the remaining 31 suspected mitochondrial disorder patients, 14 patients genetically diagnosed with mitochondrial disease, 7 patients genetically diagnosed with congenital muscular dystrophy and 2 unaffected individuals.

The 5 individuals genetically diagnosed with neurodevelopmental syndrome had pathogenic mutations within the *ATG7* gene (Table 4.2). The phenotypes of these patients is described described in a recent publication (95). RNA-sequencing data derived from this sample set was analysed to find i) differentially expressed genes common across all 5 patients, and ii) aberrantly expressed genes specific to a single family. For the former, all 5 patients were grouped together as the case cohort. As before, the control set consisted of all other individuals - 32 patients clinically diagnosed with suspected mitochondrial disorder, 14 patients genetically diagnosed with mitochondrial disease and 7 patients genetically diagnosed with congenital muscular dystrophy and 2 unaffected individuals. For the latter, each of the 5 patients was compared independently to a set of the controls. The control samples used were the same 55 individuals as described above.

4.2.2 Culturing fibroblasts

Skin biopsies were taken from each individual and fibroblast lines generated by each of the recruiting centres. After receiving frozen cell pellets, fibroblast cell lines were cultured locally in Dulbecco’s Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum and 0.05 g/ml uridine. Fibroblasts were harvested by first detaching cells using TrypLE Enzyme (Thermofisher), followed by washing with Dulbecco’s Phosphate Buffered Saline (DPBS) prior to storage at -80°C. Total RNA was extracted from fibroblast pellets using Qiagen RNeasy kit and following the manufacturer's protocol. In order to assess RNA quality, RNA integrity numbers (RIN) were measured using Agilent Technologies 2100 Bioanalyzer or Agilent 4200 Tapestation with all RIN values found to exceed 8.0

4.2.3 RNA-sequencing, alignment and quality control

Libraries for sequencing were prepared by Emil Gustavsson from the Ryten lab using the Illumina TruSeq Stranded mRNA Library Prep kit using 50 ng of total RNA in the initial reaction; fragmentation and PCR steps were undertaken as per the manufacturer's instructions. Final library concentrations were determined using a Qubit 2.0 fluorometer and pooled to a normalized input library. Pools were sequenced using the Illumina NovaSeq 6000 Sequencing system to generate 150 bp paired-end reads with an average read depth of 100 million reads per sample. Pre-alignment quality control including adapter trimming, read filtering and base correction were performed using *fastp*, an all-in-one FASTQ preprocessor (v0.20.0) (80). Reads were aligned using *STAR* 2-pass (v2.7.0) to the GRCh38 build of the reference human genome (hg38) using gene annotation from Ensembl v97 (12, 81). Novel junctions discovered in the 1st pass alignment were used as input to the 2nd pass to improve the sensitivity of junction detection. Reads were required to uniquely map to only a single position in the genome. The minimum required overhang length of an annotated and unannotated junction was set to 3 and 8 base pairs, respectively. The output BAM files underwent post-alignment QC using *RSeQC* (v2.6.4), with all samples passing quality control after manual assessment (82).

4.2.4 Detection of aberrant spliced events using *dasper*

dasper (v1.1.0) was used to detect aberrant splicing events in RNA-sequencing data derived from each of the 32 patients with suspected mitochondrial disorders. Junctions in the SJ.out format as well as BigWig files were used as input for *dasper*. SJ.out files were obtained through *STAR* during alignment, whilst BigWig files were generated from the BAM files using *megadepth* (59). *dasper* operates in a 1-vs-all manner, comparing each patient independently to a remaining set of controls. Here, each of the 32 patients were compared to a set of 54 control samples (See methods section 4.2.1).

First, the junction data was filtered to remove reads that were likely attributed to noise or technical artefacts. Specifically, junctions were required to have at least 5 counts in at least 1 sample. In addition, junctions that had an implied intron shorter or longer than any existing known intron (20-1,000,000 base pairs). Finally, junctions were required to not overlap any ENCODE blacklist regions to avoid those attributed to potential mapping errors (86). Next, junctions were classified with respect to the existing annotation (Ensembl v97) (12) into the categories "annotated", "novel acceptor", "novel donor", "novel combo", "novel exon skip", "ambig gene" and "unannotated" (Figure 3.1). "unannotated" and "ambig gene" junctions were removed, as assignment to a gene was required for downstream prioritisation. Junctions were clustered together through the sharing of an acceptor or donor site. The raw count of each junction was normalised using the total number of counts in its associated cluster.

Coverage was obtained across the 3 regions of interest, namely the intron and 2 flanking exonic regions, for each junction within each sample. Coverage was normalised by dividing the coverage across these 3 regions of interest, by the mean coverage across exonic regions of the corresponding gene. To determine the exonic regions for each gene, the MANE-select transcript was used.

The normalised counts of junction and coverage were compared between each patient of interest and the control set using the z-score approach. For a given junction, only coverage from a single region, namely that with the highest absolute z-score, was kept for input into the downstream isolation forest model. The isolation forest outlier was implemented via the python *sklearn* library (version 1.0). Junction and coverage z-scores for each patient were placed into an isolation forest model with default parameters (estimators = 100, contamination = "auto", max features = 1.0) to generate an outlier score that represented the aberrancy of each splicing events in the patient in comparison to controls (96). For each patient, splicing events were ranked upon this outlier score. The output of *dasper* is a ranked list describing the aberrancy of each splicing event within each of the patients, with a rank of 1 representing the most aberrant splicing event in that patient.

4.2.5 Generating gene count matrices

The input for the tools *OUTRIDER* and *DESeq2* is a gene count matrix. Gene count matrices are formatted with rows corresponding to genes, columns corresponding to samples and each cell containing the count for that gene-sample pair. For each of the 60 individuals who underwent RNA-sequencing, gene count matrices were obtained using RNA-SeQC (v2.3.4) with the BAM files as input (97). Ensembl v97 reference annotation was used to obtain gene definitions. Gene models were collapsed matching the protocol used in the GTEx pipeline (https://github.com/broadinstitute/gtex-pipeline). Importantly, amongst other pre-processing steps, this collapse removed overlapping intervals between genes to avoid conflation of gene counts.

4.2.6 Detection of aberrant expressed genes using *OUTRIDER*

The input for *OUTRIDER* was the gene count matrix for each patient sample. *OUTRIDER* (v1.6.1) was used to find aberrantly expressed genes in each of the patients with suspected mitochondrial disorders or an ATG7-associated neurometabolic disorder (33). A 1-vs-all experiment design was used, whereby each patient was compared to a set of controls. See methods section 4.2.1 for details of the control samples used for each analysis.

OUTRIDER was used to normalise gene counts into fragments per kilobase of transcript per million mapped reads (FPKM), the gene expression standard. FPKM matrices were filtered for genes that were expressed with an FPKM above 1 in 5% of samples. The optimal encoding dimension of the autoencoder was obtained using default *OUTRIDER* settings. During this fitting process, related samples which had the same pathogenic variant were masked to avoid obscuring outlier expression events that may have been shared across samples. Gene expression values were then corrected using the autoencoder. The default statistical test within *OUTRIDER* was applied to find significantly aberrantly expressed genes within each patient. P-values were corrected for multiple testing using the Benjamin Hochberg method and 0.05 was used as the significance threshold.

4.2.7 Disease and Mitocarta gene lists

The full set of Online Mendelian Inheritance in Man morbid (OMIM-morbid) genes were obtained using the biomaRt R package (v2.40.5) and based on Ensembl v100 (12, 98). A wider set of genes known to cause mitochondrial disorders was downloaded from the recent paper by Kremer and colleagues (24). Genes found to have strong experimental support of localisation to the mitochondria were obtained through Mitocarta 3.0 (https://www.broadinstitute.org/files/shared/metabolism/mitocarta/human.mitocarta3.0.html) (99).

4.2.8 Filtering for aberrant events capable of causing a suspected mitochondrial disorder

Patients without a genetic diagnosis (32 cases) were thought to have rare, genetic disorders with a suspected mitochondrial component (Tables A.3, A.4, A.5, A.6). Therefore, after the detection of aberrantly expressed genes and aberrant splicing events, I applied a set of automated filters to try to detect genes that were capable of causing such a phenotype (Figure 4.2). Specifically, I filtered for any event (aberrant expression or splicing) that affected genes: i) within the set of OMIM-morbid disease genes or, ii) genes known to cause mitochondrial disorders, or iii) genes that were part of Mitocarta 3.0.

For the genes that remained, I used DisGeNET to identify if any of these genes had published disease associations (100). For genes with disease associations, I manually checked these publications for evidence of association to a mitochondrial function. Additionally, I used https://pubmed.ncbi.nlm.nih.gov to search for the "Gene name + mitochondria" to find primary literature that had evidence of the genes being involved in mitochondrial function. This information was collected and manually compared for compatibility with the patient phenotype in collaboration with the recruiting clinical teams.

Aberrant splicing events were curated by predicting whether they would be expected to disrupt the protein-coding sequence. First, *dasper* was used to generate sashimi plots of each aberrant splicing event. Then, splicing events were visually inspected and classified into categories: cryptic exon, exon extension, exon truncation, intron retension, novel start site, exon skipping and change in annotated alternative splicing (isoform switching). Using this classification and the co-ordinates of the specific aberrant junctions a predicted transcript structure was built. This transcript structure was compared to the existing canonical transcript to determine the functional consequence of the change upon predicted protein sequence. When more than one annotated transcript was present for a gene, the MANE-select transcript was chosen for this comparison. In order to find the DNA-sequence of the aberrant transcript, the tool *ORFik* (v1.12.5) was used find the open reading frame (ORF). If multiple ORFs were possible, the longest ORF with the same start site as the canonical ORF was selected. Next, *Biostrings* (v2.60.1) was used to obtain the DNA and amino acid sequence for the ORF of both the aberrant and canonical transcripts (101). The amino acid sequence of the aberrant and canonical transcripts was manually compared to determine severity of the protein coding consequence. In order to highlight the functional importance of disrupted exons, conservation of these exons as well as overlapping domains were retrieved. The phastcons20 score was used to obtain the mammalian conservation of exons of interest (102). The Ensembl web browser was used to find protein domains that overlapped the transcript of interest (12).

Together, manual assessment of the information regarding the compatibility of the gene with the patient phenotype alongside the aberrant splicing consequence was used to identify candidate genes.

4.2.9 Whole-exome sequencing

Exome library preparation was performed by members of the Lily consortium using Agilent SureSelectXT All Exon V6 according to manufacturer's instructions, followed by sequencing on an Illumina HiSeq 3000/4000 with 100 bp paired-end reads. Variant calls were generated with an in-house pipeline as previously described with minor alterations. In brief, resulting reads were aligned to the reference genome (hg19) with the NovoAlign (Novocraft, 2014) alignment tool. Clonal reads resulting from polymerase chain reaction, optical errors, and reads mapping to multiple locations were discarded from further analyses. SNPs and small insertion/deletions were identified and filtered for quality with *SAMtools* (103). Variant files were annotated with respect to genes and functional consequences using the *ANNOVAR* tool. Further annotation included variant pathogenicity predictors, *CADD*, *SIFT*, and *PolyPhen2* as well as information on variant novelty and estimated population frequencies by cross-referencing identified variants with publicly available data from 1,000 genomes, ESP and ExAC datasets, and $\tilde{1}$,000 in-house control exomes processed with the pipeline described above (104–109).

4.2.10 Sanger sequencing and splice prediction

The genetic defect underlying the aberrant mRNA splicing pattern was investigated using Sanger sequencing. The characteristics of the aberrant splicing events were considered to allow the location of the genetic variant to be estimated. For example, for a cryptic exon event, sequencing analysis of the region spanning the cryptic exon donor and acceptor motifs would be likely to contain the variant of interest. Prediction of splicing changes and Sanger sequencing was performed by Dr Charlotte Alston. This approach was implemented to search for putative splicing variants within *ECHS1* and *DNAJA3*.

Confluent fibroblasts were harvested and RNA was extracted using the ReliaPrep™ RNA Miniprep System according to the manufacturer's protocol (Promega). 1ug of fibroblast-derived RNA was reverse transcribed using 0.5ug random hexamers and 200U M-MLV Reverse Transcriptase (Promega) supplemented with 25U RNasin RNAse inhibitor (Promega). PCR was performed using Go Taq Hot Start polymerase (Promega) and enzymatic clean-up of PCR products was performed using Exonuclease I (Thermo Scientific) and FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific). Sanger sequencing was performed using the BigDye terminator kit v3.1 (Life Technologies) and sequencing chromatograms were visualised using FinchTV (v1.4.0) and aligned against the Human Genome (GRCh38) using BLAT (110).

DNA sequence for the region of interest was obtained using the 'View Data' tool within the UCSC genome browser; oligonucleotide primers were designed using Primer3 and manufactured by Integrated DNA Technologies. PCR amplification of 50ng genomic DNA was achieved using Go Taq Hot Start polymerase (Promega) using the following Thermocycling conditions (on ABI9700 machine): 95°C – 2mins; 30 cycles of [95°C – 1min, 62°C – 1min, 72°C – 1min] and a final extension at 72°C – 10mins. The primer sequences for *ECHS1* and *DNAJA3* are found in Table 4.1.

	Gene	Strand	Sequence
1	*ECHS1*	Forward	5'-CCAAGAAATAACTGGCGGGT-3'
2	*DNAJA3*	Forward	5'-TGTTTGGATCCATTGCCCAC-3'
3	*ECHS1*	Reverse	5'-AGTGATTGACAAGGTGAAGCA-3'
4	*DNAJA3*	Reverse	5'-CACACGCGCCTATAAACACA-3'

Table 4.1 Forward and reverse primer sequences for *ECHS1* and *DNAJA3*.

4.2.11 Variant classification (ACGS 2020 guidelines)

Variants were classified using the standardised American College of Medical Genetics (ACMG) guidelines which assign weight using a variety of criteria, including population frequency data, in silico prediction tools, previous/additional cases and functional study data. A culminative score is assigned depending on the final 'weight' of evidence, either class 5 (pathogenic), class 4 (likely pathogenic), class 2 (likely benign) and class 1 (benign) whilst variants remaining of uncertain pathogenicity (VUS) are given class 3 status.

4.2.12 Detection of differentially expressed genes and pathways

DESeq2 (v1.26.0) and *gprofiler2* was applied to the 5 the patients with ATG7-associated neurometabolic disorder to find differentially expressed genes and pathways on comparison to control samples (consisting of 55 individuals, details in the methods section 4.2.1). (31, 111). The input for *DESeq2* were the gene count matrices. Gene count matrixes were filtered for genes with an FPKM above 1 in 5% of samples. *DESeq2* was applied with default settings using 5 patients grouped as the case cohort. The Benjamin Hochberg method was used for multiple test correction. Genes with a corrected p-value less than 0.05 and an absolute log-fold change greater than 1.5 were considered to be differentially expressed. *gprofiler2* was then applied to the differentially expressed genes to find pathways that were dysregulated across the 5 patients. Inputted genes were ranked by their p-value. All expressed genes within the expression-filtered gene count matrixes were used as the background list.

4.3 Results

4.3.1 Patient cohorts used in this study

The patients analysed within this chapter originated from two cohorts; 32 patients that were diagnosed with a suspected mitochondrial disorder and 5 patients that were diagnosed with an ATG7-associated neurometabolic disorder.

All 32 patients with a suspected mitochondrial disorder were recruited from UK-based institutions. The origin, phenotype, demographic and diagnostic criteria of the 32 patients is detailed in the tables A.3, A.4, A.5, A.6. Dependent on the diagnostic routine that patients had undergone, they fell into 2 categories; those which had undergone WGS, those which had undergone WES and those which had panel-based sequencing (Table A.6). Irrespective of DNA sequencing method, none of the patients had a genetic diagnosis. It was expected that RNA-sequencing could provide diagnostic utility for this cohort through the detection of aberrant expression and splicing events.

Samples from the cohort of 5 patients with a neurometabolic disorder were analysed as part of a collaboration with Professor Robert Taylor. These 5 patients originated from 4 unrelated families and were all diagnosed with pathogenic variants within the gene *ATG7* (95). All patients were known to have biallelic missense or acceptor splice site variants (Table 4.2). Despite the pathogenicity of these novel variants being confirmed through detailed functional testing including complementation studies in human and yeast models, it was hypothesised that RNA-sequencing would help better inform our understanding of the pathogenesis of this new congenital disorder of macroautophagy through detection of dysregulated genes and pathways (95).

	Sample ID	Family	Gene Symbol	Variant type	Variant HGVS
1	M1856.17	1	ATG7	missense;acceptor	c.1975C>T;c.2080-2A>G
2	M0920.18	2	ATG7	missense;missense	c.1727G>A;c.1870C>T
3	M0921.18	2	ATG7	missense;missense	c.1727G>A;c.1870C>T
4	M1111.19	3	ATG7	missense;missense	c.700C>A;c.1762G>A
5	M1716.19	4	ATG7	missense;missense	c.782A>G;c.1532G>A

Table 4.2 Pathogenic variants of 5 neurometabolic disease patients. All 5 patients were genetically diagnosed with causative variants within *ATG7*. The patients originated from 4 unrelated families and each family harboured distinct variants. The table describes the HGVS identifiers and locations of these variants.

4.3.2 Prioritisation of candidate genes through the detection of aberrantly expressed and spliced genes

RNA-sequencing has previously been demonstrated to serve as a promising diagnostic tool through the transcriptome-wide detection of aberrant RNA-level expression and splicing, complementing DNA-sequencing and allowing for the re-interpretation of variants (23, 24, 27, 28). Following this approach, I applied the tools *OUTRIDER* and *dasper* to detect aberrantly expressed genes and aberrant splicing events within the RNA-sequencing data derived from the 32 patients who remained unsolved following WES or panel-based analysis (Table A.6) (33, 112). Both *OUTRIDER* and *dasper* operate through a 1-vs-all experimental design, comparing the reads associated with genes or splicing events within each patient to a set of controls (See detailed Methods). However, the two methods differ in their approach for identifying aberrant events. *OUTRIDER* uses an autoencoder to correct for covariates in the RNA-sequencing data, then applies a statistical test to find genes with significant aberrant expression. Whilst *dasper* uses an isolation forest to rank every splicing event by its aberrancy, with a rank of 1 specifying the most aberrant splicing event in each patient. Across the 32 patients, *OUTRIDER* detected 26 significant aberrantly expressed genes, comprising of 14 up-regulated genes and 12 down-regulated genes. When applying *dasper*, I considered the top 100 most aberrant splicing events in each patient to be taken forward for further prioritisation, resulting in 3200 aberrant splicing events across 1712 unique genes. Between the *OUTRIDER* and *dasper* results, 6 genes overlapped and were found to be both expressed and spliced aberrantly. In total, I detected 3226 aberrant RNA-level events across the 32 patients, which were associated with 1732 unique genes (Figure 4.2).

In order to filter these 3326 aberrant events to set of candidate genes, I combined the RNA-sequencing results with phenotypic information for the 32 patients. All patients within the analysed cohort were clinically diagnosed with a rare, genetic disorder which was found to cause a degree of mitochondrial dysfunction. Therefore, I filtered for aberrant events which were associated with genes that were known to cause Mendelian disease or had evidence of localisation to the mitochondria. Namely, this included 4183 genes that were previously reported to cause an OMIM-morbid disease, an additional wider set of 283 genes reported to cause mitochondrial disease by Kremer and colleagues and 1136 genes that were demonstrated to localise to the mitochondria through the Mitocarta project (24, 98, 99). After these filters, 5 aberrantly expressed genes and 619 aberrant splicing events were retained, which were associated with 391 unique genes (Figure 4.2). Of these 391 genes, 7 (2%) were already known to cause mitochondrial disease, whilst 46 (12%) were solely part of Mitocarta and had no known association with any disease. The majority of genes were part of the OMIM-morbid catalogue (344/88%). There was a degree of overlap between the gene lists, with 5 (1%) genes being part of all 3 lists and 16 (4%) genes being both an OMIM-morbid gene and part of Mitocarta (Figure 4.1). Together, this automated gene-level filter was able to reduce the set of 3326 aberrant events to 624 (Figure 4.2).

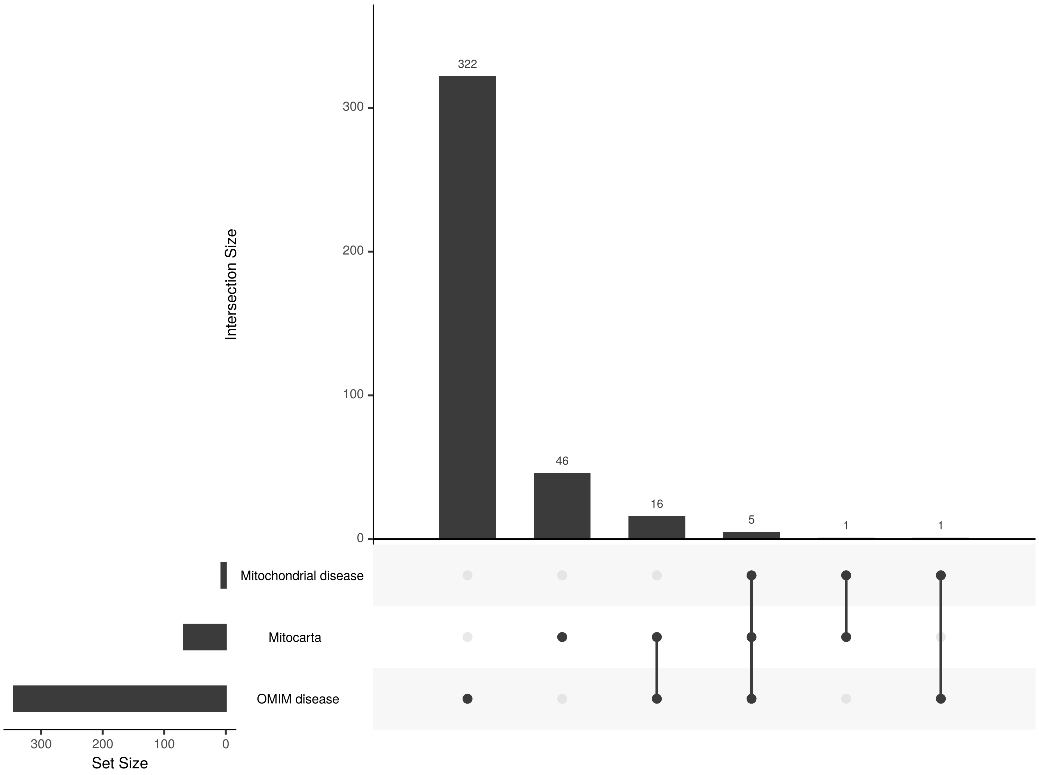


Figure 4.1 Intersection between aberrant events and mitochondrial disease, OMIM-morbid and Mitocarta gene lists. Aberrantly expressed genes and aberrant splicing events were identified using the RNA-sequencing data. The genes associated to these aberrant events were filtered for only those that were associated with mitochondrial disease genes, OMIM-morbid genes and genes which were part of Mitocarta. Of the remaining aberrant events, the plot displays the number of genes that were present within a single gene list, spanned across two gene lists or were present across all three.

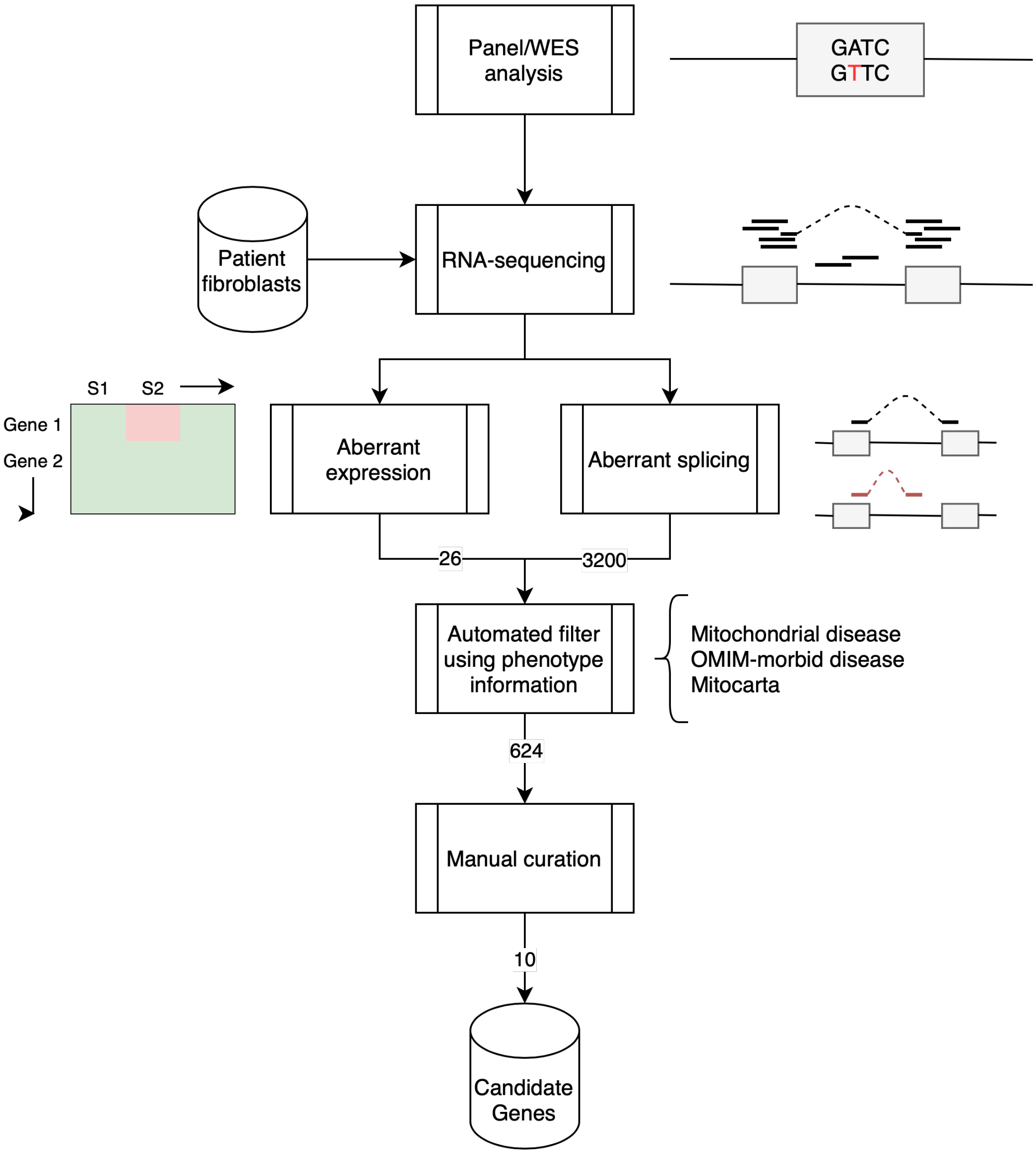


Figure 4.2 Prioritisation of candidate genes using RNA-sequencing. The workflow describes a set of analyses and filters that were used to identify candidate genes from patient RNA-sequencing data. RNA-sequencing was performed on fibroblast cell lines derived from 32 patients who remained unsolved after WES or panel-based analysis. Next, the tools *OUTRIDER* and *dasper* were applied to the RNA-sequencing data and identified 26 aberrantly expressed genes and 3200 aberrant splicing events across the 32 patients respectively. The resulting aberrant events were filtered for those that had evidence of being able to cause the patients' phenotypes, reducing this list to 624 aberrant events. Finally, these remaining examples were manually curated, which involved further investigation of prioritised gene function and predicting the functional consequence of the aberrant splicing events. Together, this approach discovered 10 candidate genes.

Finally, I manually curated the 624 prioritised aberrant events to find the most promising candidate genes. In brief, I searched the previous literature for consistency between the prioritised gene function or associated disease and the patient's reported phenotype. Specifically for aberrant splicing events, I also investigated whether the splicing change would be predicted to cause a functional disruption with respect to the canonical, MANE-select transcript. In particular, I predicted whether the aberrant splicing event would induce a frameshift and/or stop codon in the coding sequence, or cause a reduction in expression of the major protein-coding transcript. Overall, I discovered candidate genes for 10 (31%) out of the 32 patients (Table 4.3). This included 10 genes that were part of the OMIM-morbid catalogue, 5 genes associated with mitochondrial disease and 1 gene that was solely part of Mitocarta and not previously known to cause disease. Thus, I demonstrated that using automated prioritisation and manual curation of aberrant events detected through RNA-sequencing, I was able to discover candidate genes for a third of the unsolved cases recruited in this study.

	Gene Symbol	Aberrant Event	Mitochondrial disease	Mitocarta	MIM number
1	ECHS1	expression/splicing	TRUE	TRUE	616277
2	TBCK	expression/splicing	FALSE	FALSE	616900
3	DNAJA3	expression/splicing	FALSE	TRUE	
4	SURF1	splicing	TRUE	TRUE	256000, 616684
5	TPK1	splicing	TRUE	FALSE	614458
6	CHCHD10	splicing	TRUE	TRUE	615048, 615911, 616209
7	NSUN3	splicing	TRUE	TRUE	
8	FANCD2	splicing	FALSE	FALSE	227646
9	ACE	splicing	FALSE	FALSE	267430, 612624, 614519
10	GLDC	splicing	FALSE	TRUE	605899

Table 4.3 Candidate genes discovered through RNA-sequencing. Details of the 10 candidate genes that were discovered through RNA-sequencing of the 32 unsolved cases. Of the 10 candidate genes, 3 were found to had evidence of being both aberrantly expressed and spliced from the RNA-sequencing data, whilst the remaining 9 were only found to be aberrantly spliced.

4.3.3 Diagnosis of unsolved patients using RNA-sequencing

The 10 candidate genes identified on the basis of aberrant splicing or expression were further investigated in collaboration with the teams involved in their care. Here, I highlight two examples of potential diagnoses made through the follow up of the candidates *ECHS1* and *DNAJA3*.

From the RNA-sequencing data, I discovered that *ECHS1* is both aberrantly expressed (p-value: 9x10^{-7}) and aberrantly spliced (rank: 19) in the patient M2566.15. Previous literature has demonstrated that disruptions of *ECHS1* function are known to cause mitochondrial disease and Leigh syndrome (24, 98, 113). Furthermore, the observed decrease in expression (z-score: -7.3) of *ECHS1* is consistent with previous reports that deficiency of *ECHS1* can cause mitochondrial encephalopathy with cardiac involvement (114). The consequence of the aberrant splicing is observed to be a cryptic exon insertion event represented in the RNA-sequencing data by the appearance of 2 novel junction events, one with a novel acceptor and the other with a novel donor site (Figure 4.3). This inserts a 191bp cryptic exon between exon 4 and exon 5 of the ENST00000368547 MANE-select transcript(NM_004092.4) and predicts a r.514_515ins515-

753_515-563 at the mRNA level and p.Gly172Glufs*16 at the predicted protein level. Therefore, it is expected to induce a premature stop codon, which truncates exons 5-8 of the *ECHS1* protein. Although it is likely that this aberrant transcript would undergo NMD, given that I was able to detect the splicing disruptions through RNA-seq, I also checked the functional consequence of the truncation. I found that mutations within exon 5 have been shown to cause Leigh syndrome and additionally, exons 5-7 are highly conserved (phastCons20: 0.78, 0.69, 0.75), suggesting that they are functionally important (115, 116). Sanger sequencing of the regions flanking the cryptic exon revealed a rare homozygous variant chr10:g.133367556T>C (hg38) (NM_004092.4(*ECHS1*): c.515-563A>G), which is completely absent in the gnomAD database (117). Furthermore, this variant was predicted by multiple splicing prediction tools to cause the cryptic exon event observed in the RNA-sequencing data (Figure 4.4). Finally, the patient's phenotype (neonatal onset, persistent lactic acidosis, encephalopathy and myopathy), was consistent with other *ECHS1* cases in the literature (118, 119). Taken together, this example shows how RNA-sequencing data can be used to highlight decreased expression and presence of a cryptic exon event within *ECHS1*, leading to detection of the rare, non-coding variant as well as providing the evidence for its pathogenicity.

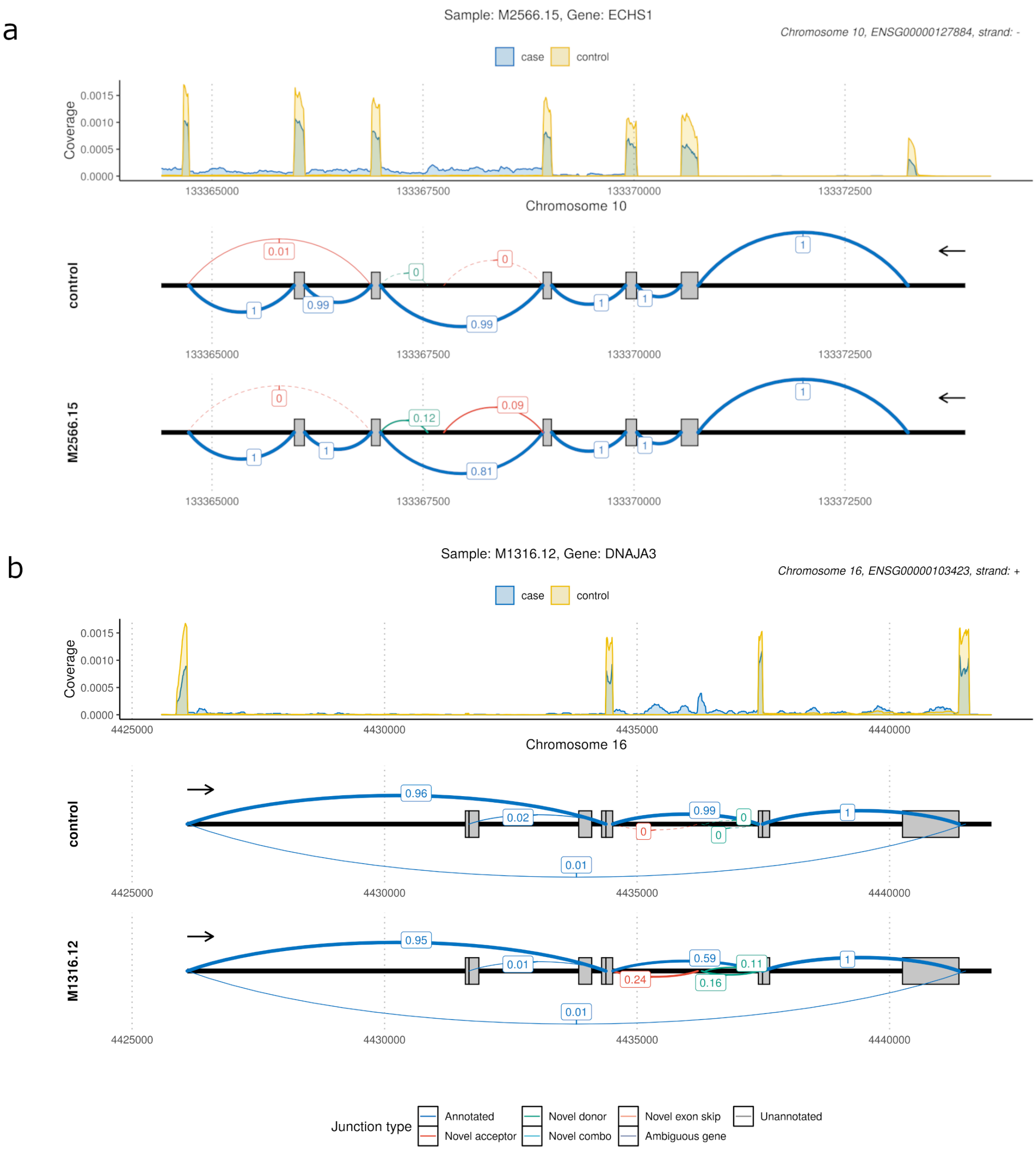


Figure 4.3 Detection of cryptic exon insertion events within *ECHS1* and *DNAJA3*. Sashimi plots illustrate the aberrant splicing events found in the candidate genes *ECHS1* and *DNAJA3*. Both plots detail the coverage and junctions across the region of the transcripts which contain the aberrant splicing events. The top track represents the coverage across the region, whereas the bottom tracks display the junctions. Within the junction track, grey blocks highlight the position of the annotated exons and curved lines represent the junctions. Junctions are labelled with their corresponding normalised count found in the case sample or averaged across the control samples. The arrow represents the direction of transcription for the plotted gene. For both *ECHS1* and *DNAJA3*, the cryptic exon insertions are characterised by non-overlapping novel acceptor (red) and novel donor (green) junctions within the same intron, alongside an increase in coverage across the associated region. For *DNAJA3*, the presence of 2 novel donor junctions suggests that 2 distinct cryptic exons are created within this gene.

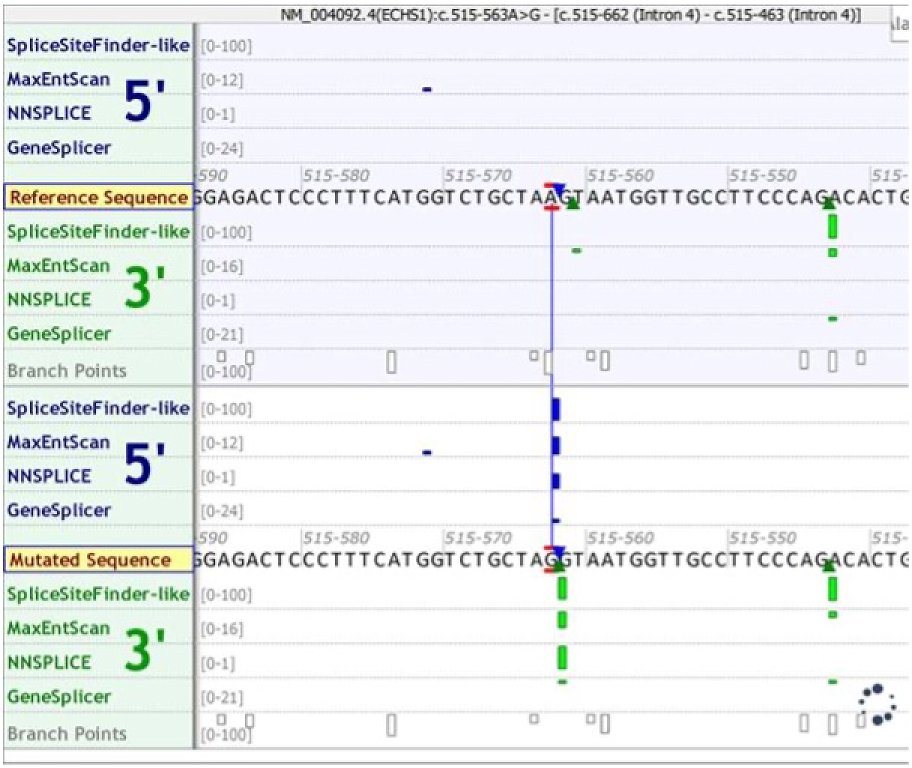


Figure 4.4 Prediction of *ECHS1* cryptic splice sites from DNA sequence. The wild-type sequence is depicted in the top panel, whilst the patient's DNA sequence is displayed on the bottom. A variety of in silico tools were used to predict the splice sites from the DNA sequences using the Sofia Genetics' Alamut Visual Software (v2.1.2). Donor changes are represented by the blue arrows (above the DNA sequence) and acceptor changes by the green arrow (below the DNA sequence).

As in the case of M2566.15, RNA-sequencing data from the patient M1316.12, identified the significant down-regulation of a specific gene, *DNAJA3*, in comparison to controls (p-value: 0.0001, z-score: -6.44) as well as evidence of aberrant splicing in the same gene (rank: 13). Dysfunction of *DNAJA3* has not previously been reported to cause disease, however there is evidence of its mitochondrial localisation through the Mitocarta database (99). Furthermore, previous studies in mice suggest that this gene plays a crucial role in mitochondrial biogenesis. Deficiency of *Dnaja3* has been demonstrated to be lethal in mice within 10 weeks (120). The aberrant splicing event in *DNAJA3* is characterized by the appearance of 2 novel donor junctions and a novel acceptor junction (Figure 4.3). This is predicted to cause two distinct cryptic exons to be inserted into the ENST00000262375 (MANE-select v0.93) transcript structure. The existence of these two cryptic exons is supported by increased coverage across the putative exonic region. These two cryptic exons are of length 76 and 156 respectively and lie between exon 2 and exon 3. They are both predicted to induce a premature stop codon truncating exons 3-12 from the DNAJA3 protein. The disrupted exons are highly conserved (phastCons20; mean: 0.83, median: 0.90, min: 0.25, max: 0.99) and the protein regions they encode contain multiple protein domains suggesting that they are functionally important (Table A.2) (115, 121, 122). Overall, *DNAJA3* represents a promising candidate and a possible novel gene-disease association discovered through RNA-sequencing. Investigation of the patient's WES data alongside functional studies are required to confirm *DNAJA3* as genetic cause of the patient's disease and this work is on-going.

4.3.4 The representation of splicing disruptions in RNA-seq

Abnormalities in splicing can impact on transcript structures in multiple ways and these include: exon skipping, cryptic exon insertions, exon extensions or truncations, intron retentions and changes in the usage of annotated transcripts (23). An improved understanding of the specific form of splicing disruption can be used to highlight the region containing the causative variant, as well as provide the framework for developing therapeutics that modulate splicing (44). With this in mind, I investigated the types of aberrant splicing events detected within the 10 candidate genes. Overall, I detected 1 exon skipping event, 1 novel start site, 2 cryptic exon events, 2 isoform switching and 4 exon extension events (Table 4.4). Here, I describe the representation of these splicing disruptions in the RNA-sequencing data through the examples *GLDC*, *TPK1*, *ECHS1*, *DNAJA3*, *SURF1* and *TBCK* (Figure 4.5, 4.7 and 4.6).

	Gene Symbol	Splicing event	DNA consequence	Protein consequence
1	TBCK	isoform switch	-	reduction in protein expression
2	CHCHD10	isoform switch	-	reduction in protein expression
3	ECHS1	cryptic exon	premature stop codon	truncation
4	DNAJA3	cryptic exon	premature stop codon	truncation
5	SURF1	exon extension	frameshift	downstream exon AA sequence
6	NSUN3	exon extension	premature stop codon	truncation
7	FANCD2	exon extension	premature stop codon	truncation
8	TPK1	exon skipping	premature stop codon	truncation
9	ACE	exon skipping, exon extension	premature stop codon	truncation
10	GLDC	novel start site	novel start site	truncation of upstream exons

Table 4.4 Types of aberrant splice events detected across the 10 candidate genes. All 10 candidate genes that were prioritised after manual curation had evidence of being aberrantly spliced according to the RNA-sequencing data. The gene symbols, type of splicing event and consequence to the DNA/protein sequence are described here.

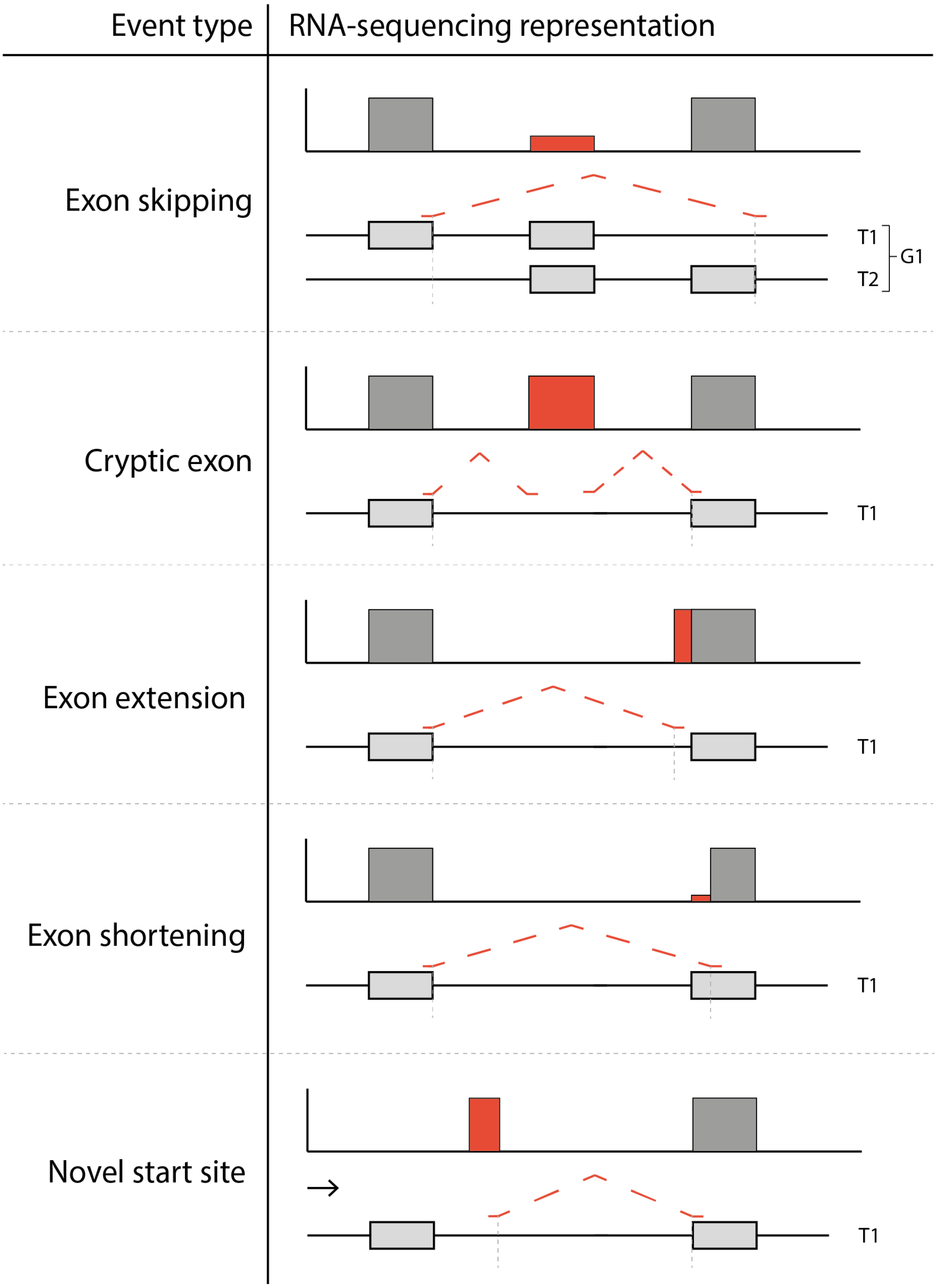


Figure 4.5 Representation of aberrant splicing events in RNA-sequencing data. The schematic illustrates how the 5 aberrant splicing types are represented in the RNA-sequencing data in terms of junctions and coverage. Top and bottom panels for each type of aberrant splicing event detail the coverage and junctions respectively. On the coverage panel, grey blocks indicate the coverage across annotated exons and red blocks indicate coverage associated with the novel splicing event. On the junction panel, red dashed lines show the expected junction location and grey blocks indicate the annotated exons.

Cryptic exons are represented in the RNA-sequencing data through two, non-overlapping novel donor and novel acceptor junctions that lie within the same intron. Furthermore, depending on their level of expression and noise across the associated intron, cryptic exons may be supported by an increase in coverage across the exonic region, as observed in *DNAJA3* (Figure 4.3). In the example of both *ECHS1* and *DNAJA3*, this would be expected to lead to disruption of the protein-coding sequence through the introduction of a premature stop codon, which truncated all exons downstream of the aberrant event. Furthermore, given the drop in expression, it is likely that the aberrant transcripts generated by these cryptic exon events within *ECHS1* and *DNAJA3* undergo nonsense-mediated decay. For *ECHS1*, as pathogenic variants which create cryptic exon insertions are likely to occur within or in close proximity to the novel splice sites, this facilitated the primer design and sanger sequencing, which consequently detected the causative variant in this patient.

Novel exon extension and truncation events are characterised in the RNA-sequencing data by a single novel acceptor or novel donor junction as well as a shift in coverage that supports the novel exon boundary in the case vs the control. Depending on whether the novel splice site falls within an intron or an exon, this will result in an extension or truncation of an exon respectively. For example, I detected an exon extension event in the gene *SURF1*. Here, exon 7 of ENST00000371974 is extended by 31bp which is predicted to cause a frameshift affecting the downstream exons 8 and 9 (Figure 4.6). Additionally, a shift in coverage is also observed, supporting the position of the novel exon boundary. Variants that disrupt the annotated splice site, create a novel splice site, or favour the usage of usually dormant splice sites can generate exon extension events. Such variants would be expected to fall near or within the novel or annotated splice sites.

Novel start site events are also represented by the appearance of a single novel acceptor or novel donor junction. However, the increase in coverage between the novel splice site and the annotated exon boundary is not contiguous, which distinguishes novel start site events from extension or truncation of exons. For example, in *GLDC* a novel donor junction is observed with a novel end that falls deep within the intron. The coverage supporting this novel event does not connect to the upstream annotated exon and there is an absence in coverage across all upstream exons. Together, this suggests the usage of a novel start site (Figure 4.7). This is predicted to generate an aberrant transcript beginning at the novel start site and containing the downstream exons (Figure 4.7). With respect to the MANE-select transcript ENST00000321612, this will truncate the exons 1-13. Similar to exon extension events, variants causing such events could lie in the vicinity of the novel or annotated splice sites.

In the RNA-sequencing data, exon skipping events are represented by the appearance of a novel junction which overlaps an annotated exon. Both ends of the novel junction must fall on annotated splice sites but their combined usage is not found within existing annotation. Depending on the expression of the aberrant event, this is likely to cause a decrease in the coverage across the skipped exon. In the example of *TPK1*, exon 4 is skipped in the MANE-select transcript ENST00000360057 (Figure 4.7). This causes a frameshift in exon 5, which as a result induces a premature stop codon truncating exons 6-8 from the TPK1 protein. Variants that cause exon skipping events may be expected to fall within or near the splice sites of the skipped exons.

A shift in the alternative splicing of known transcripts can occur to a degree to which is considered aberrant. These events do not include the appearance of a novel junction but instead are represented through a change in the expression of annotated junctions in the RNA-sequencing data. For instance, in *TBCK*, I observed a reduction in the expression of a junction that is annotated within the MANE-select transcript ENST00000394708 (Control count: 0.85, Patient count: 0.24). This is replaced in the patient with the expression of a junction that is only present in the nonsense-mediated decay transcript ENST00000467183 (Control count: 0.08, Patient count 0.72). The magnitude of this switch is large enough to be detected as aberrant using dasper (rank: 85). Interestingly, in the same patient I observe that the overall expression of *TBCK* is significantly decreased (p-value: 0.0097, z-score: -6.01), which suggests the switch to predominant expression of the nonsense-mediated decay transcript is reducing the overall expression of this gene. Variants that cause a shift in alternative splicing can fall within or near either of the disrupted, annotated splice sites.

Overall, I demonstrate the utility of RNA-sequencing to characterise splicing disruptions, a process which can be informative for determining and therefore, sequencing the likely genomic location of the causative variant.

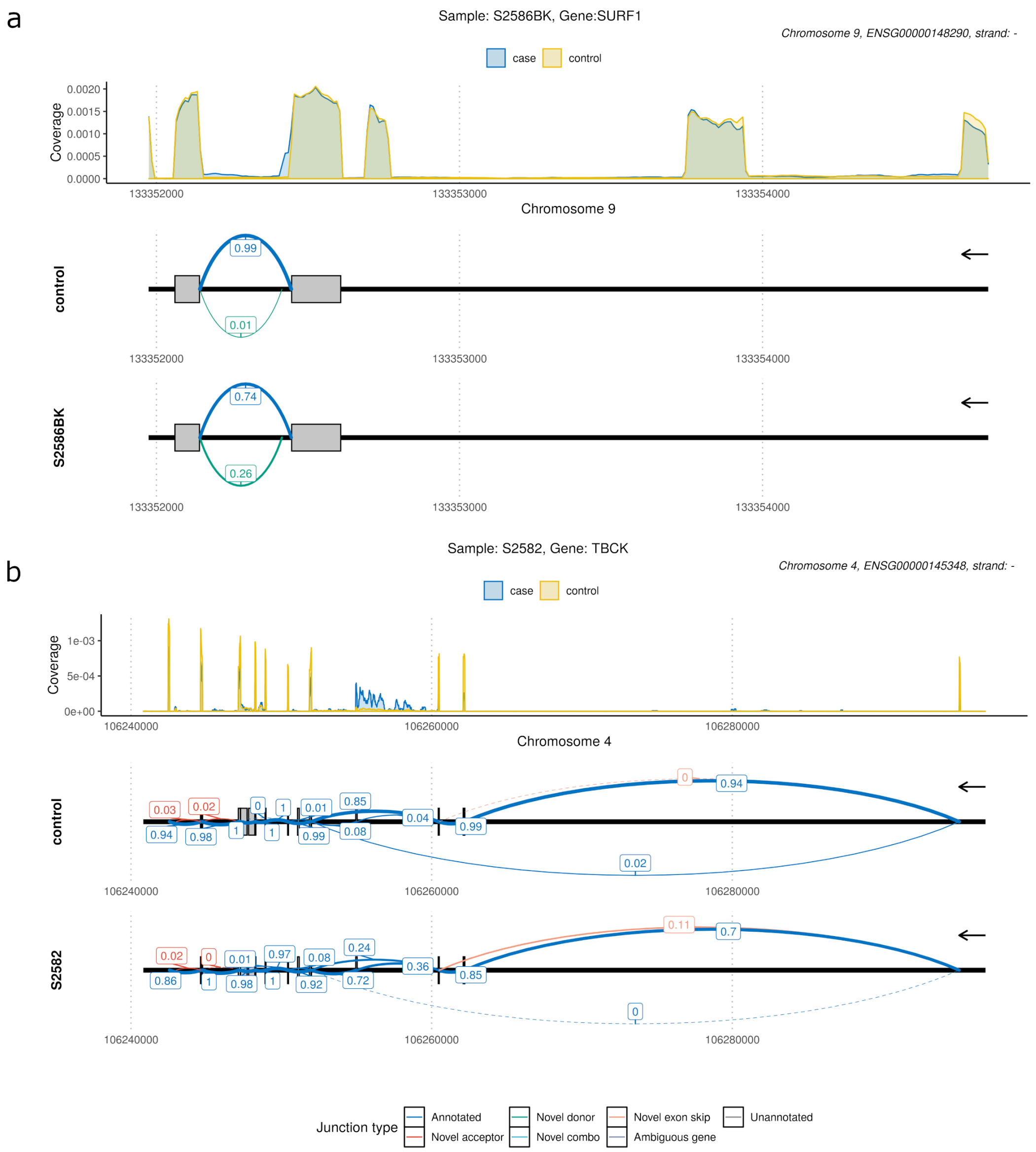


Figure 4.6 Detection of exon extension event in *SURF1* and a shift in alternative splicing of *TBCK* Sashimi plots illustrate the aberrant splicing events found in the candidate genes *SURF1* and *TBCK*. See Figure 4.3 for details of sashimi plot elements. a) In *SURF1*, a exon extension event is detected, which is represented through a novel donor junction (green) together with a shift in the coverage that matches the novel exon boundary. b) In *TBCK*, there is a shift in the alternative splicing from the canonical MANE-select transcript (Control count: 0.85, Patient count: 0.24) to a nonsense-mediated decay transcript (Control count: 0.08, Patient count 0.72). This is represented by changes in the count of annotated (blue) junctions and lacks the appearance of a novel junction.

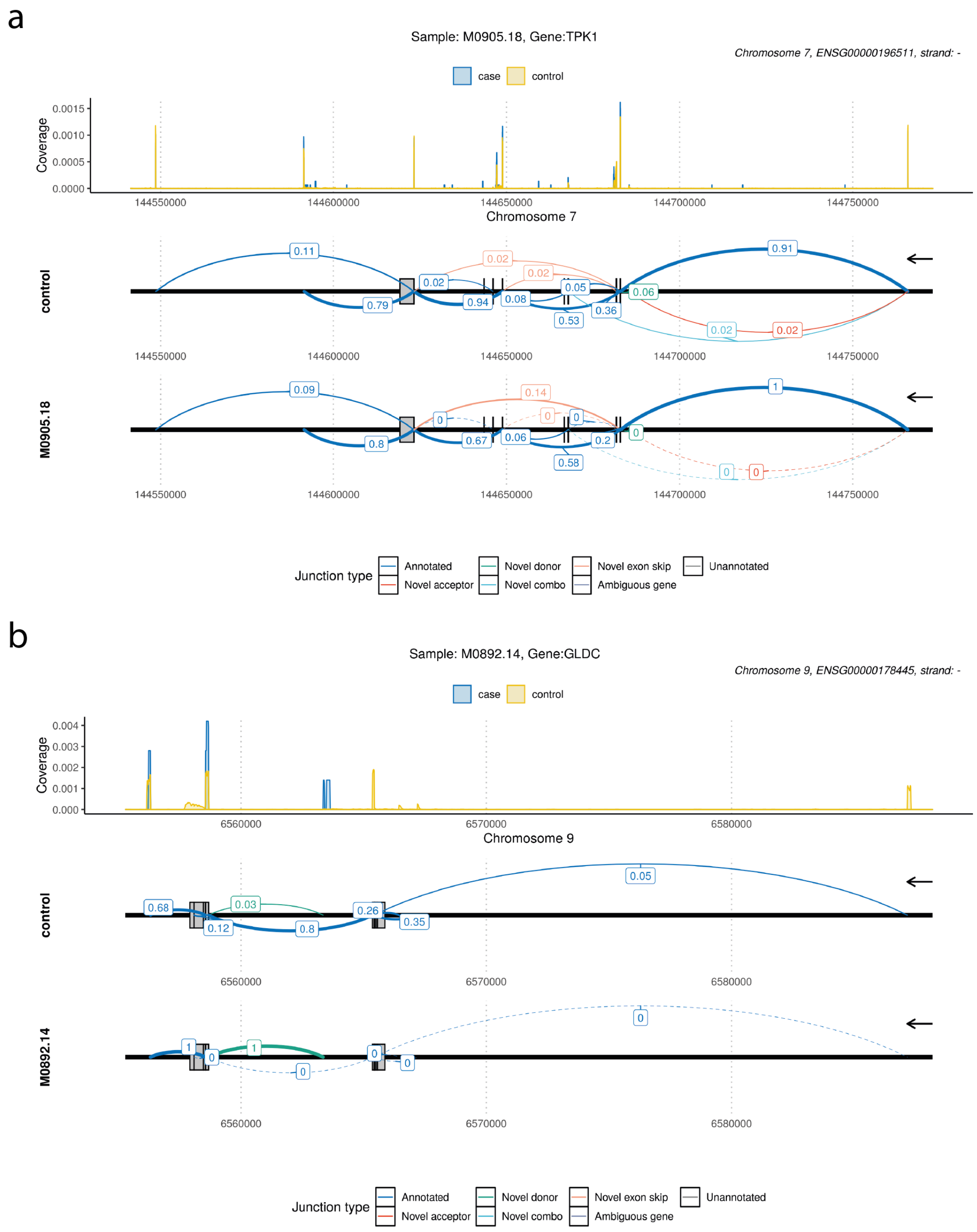
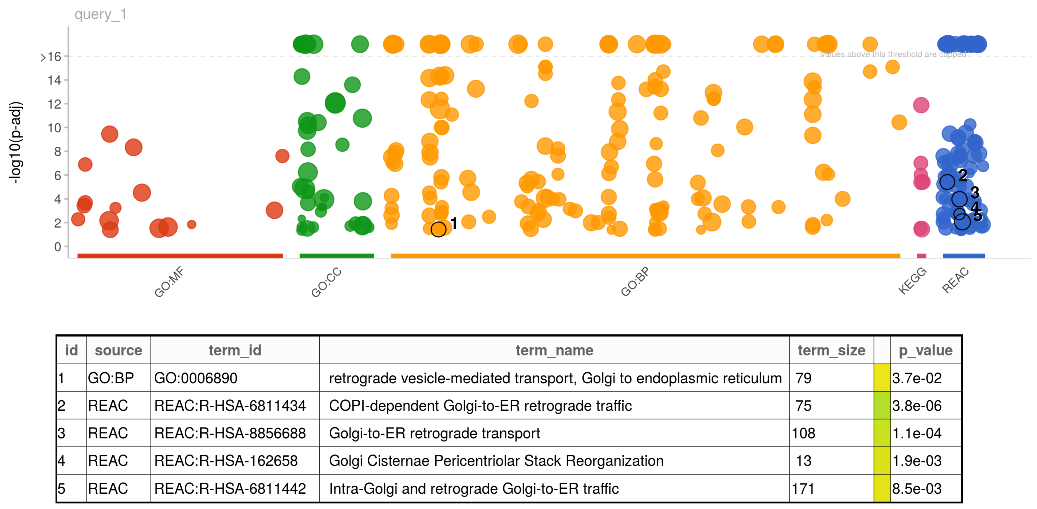


Figure 4.7 Detection of exon skipping event in *TPK1* and a novel start site in *GLDC* Sashimi plots illustrate the aberrant splicing events found in the candidate genes *TPK1* and *GLDC*. See Figure 4.3 for details of sashimi plot elements. a) In *TPK1*, a novel exon skip junction (orange) can be observed, which skips exon 4 of the MANE-select transcript ENST00000360057. This is predicted to induce a premature stop codon truncating downstream exons 6-8 from the resulting protein. b) In *GLDC*, a novel exon that contains a novel start site is detected through the changes in both junctions and coverage. A novel donor junction (green) represents the usage of a novel splice site that lies deep within intron. Alongside, a block of coverage that extends from the novel donor site and ends within the intron represents the novel exon boundaries.

4.3.5 Detection of disrupted downstream pathways and potential disease modifiers using RNA-sequencing

RNA-sequencing generates a functional readout of cellular activity, which enables the transcriptome-wide detection of disruptions to pathways and genes that occur alongside or downstream of a pathogenic event. This improved understanding of the pathogenesis of a patient's disease can provide prognostic information and permit better management of disease symptoms. Here, I applied the tools *DESeq2* and *gprofiler* to the RNA-seq data derived from a cohort of 5 patients diagnosed with neurometabolic syndrome to discover pathways that were disrupted downstream of the causative mutation (31, 111).

The cohort of neurometabolic syndrome patients consisted of 5 individuals genetically diagnosed with pathogenic variants within the gene *ATG7*, which encodes one of the core proteins involved in the autophagy pathway. (Table 4.2) (95). Given the common genetic origin of the disease within this cohort, I hypothesised that there would be common downstream pathways that were disrupted across all 5 patients. In order to investigate this, the 5 patients were grouped together as cases. Then, *DESeq2* was applied to detect genes that were differentially expressed across the cases compared to set of controls (See detailed Methods). I detected 358 genes that were significantly differentially expressed with a log-fold change of >1.5 across the 5 patients. Next, *gprofiler* was applied to find biological pathways that these 358 genes were enriched within. In general, I discovered that the majority of disrupted pathways were associated with mitotic and cell cycle functions. In addition, 5 Golgi-related pathways were also found amongst the dysregulated pathways (Figure 4.8). Consistent with this finding, recent reports demonstrate that the golgi apparatus is involved in the formation of autophagosomes and disruptions to the formation of autophagy can impact on golgi function (123).



id	source	term_id	term_name	term_size		p_value
1	GO:BP	GO:0006890	retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum	79	🟨	3.7e-02
2	REAC	REAC:R-HSA-6811434	COPI-dependent Golgi-to-ER retrograde traffic	75	🟩	3.8e-06
3	REAC	REAC:R-HSA-8856688	Golgi-to-ER retrograde transport	108	🟩	1.1e-04
4	REAC	REAC:R-HSA-162658	Golgi Cisternae Pericentriolar Stack Reorganization	13	🟨	1.9e-03
5	REAC	REAC:R-HSA-6811442	Intra-Golgi and retrograde Golgi-to-ER traffic	171	🟨	8.5e-03

Figure 4.8 Golgi-related pathways are disrupted in the *ATG7* patients Each point represents a pathway that the differentially expressed genes are enriched within. The numbered points represent the 5 pathways related to golgi function. The table below gives further detail of these 5 pathways including their full pathway name ("term_name") and "p_value".

I recognised that the *ATG7* patients had variable phenotypes and in particular, the neurological phenotype of family 2 was found to be more severe than those of the remaining families (95). I hypothesised whether this variability may be explained through disruptions outside of the *ATG7* locus. For this reason, I applied *OUTRIDER* to find genes that were specifically dysregulated within each family. In total, I discovered 264 unique genes that were aberrantly expressed across the 5 patients. Amongst these 264 genes, *VPS41*, *HMG20A* and *TBC1D3L* were found to be down-regulated in both patients from family 2, however expressed at normal levels in the remaining families (Figure 4.9). Of these 3 genes, *TBC1D3L* and *VPS41* had known functions related to the autophagy pathway. *TBC1D3L* encodes a GTPase activating protein for *RAB5*, which has a role in regulating autophagy (124). Whilst *VPS41* is suggested to confer a neuroprotective effect in neurodegenerative diseases acting via the autophagy pathway (125). For these 3 genes, the WES data of family 2 individuals was analysed for variants driving these changes in expression. This revealed a heterozygous variant that overlapped *VPS41* in the individuals from family 2. A recent report suggests that biallelic *VPS41* variants can cause cerebellar and corpus callosum defects, which aligns with the phenotypes of the patients within this family (125). However, it remains unclear whether this change in transcript abundance is a cause or result of the neurological symptoms of this family. Further investigations such as mutation burden analyses would be helpful to explore the potential of *VPS41* to operate as a genetic modifier.

Overall, I have demonstrated the capability of RNA-sequencing to detect the downstream disrupted pathways that are common across a disease as well as dysregulated genes specific to individual patients.

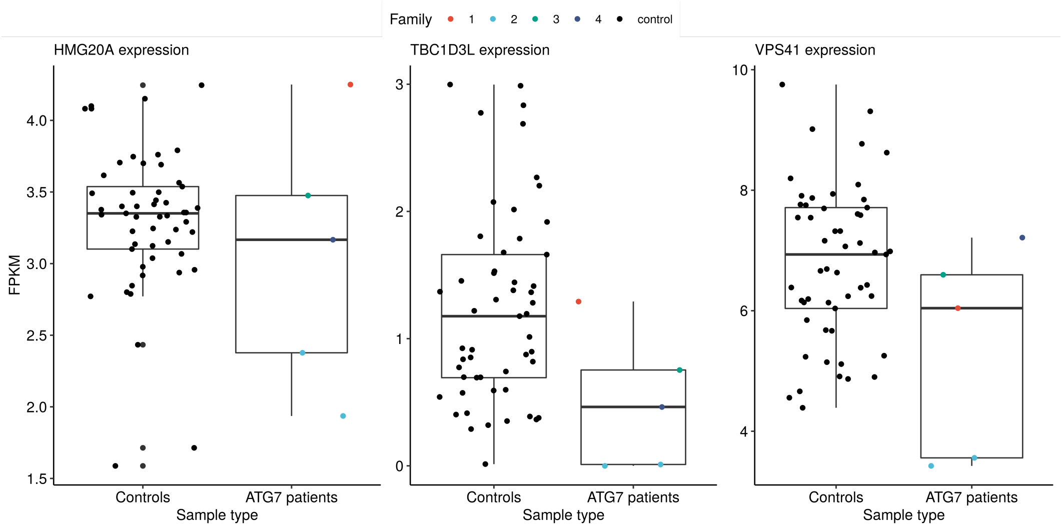


Figure 4.9 ***HMG20A*****,** ***TBC1D3L*** **and** ***VPS41*** **are expression outliers common to individuals from family 2.** Expression plots describe FRKM of the genes *HMG20A*, *TBC1D3L* and *VPS41* in controls, compared to the *ATG7* patients. Expression values in each patient are coloured, with cyan representing family 2. Individuals from family 2 have the consistently low expression for these 3 genes. After autoencoder correction of FPKM values, the expression of all 3 genes in family 2 individuals are detected to have aberrantly low expression according to *OUTRIDER*.

4.4 Discussion

In this study, I demonstrate the power of RNA-seq data as applied to patient-derived samples to enable an increase in diagnostic yield beyond conventional genetic testing, identify additional variants of relevance outside the primary locus, and improve our understanding of the pathophysiology of disease even with small case numbers. This is achieved through the application of a diagnostic workflow that integrates the assessment of aberrant gene expression with aberrant splicing. I apply this workflow to fibroblast RNA-seq data derived from 32 patients suspected of having mitochondrial disease. As a result of this analysis, a genetic diagnosis is confirmed for 1 patient and candidate genes are found for a remaining third which are still undergoing further investigation. Through specific cases, I explore the likely pathogenic mechanisms of several of these candidates. Furthermore, I use RNA-seq to improve the understanding of disease pathogenesis in a cohort of 5 patients with *ATG7*-associated disease. Not only was RNA-seq able to detect global pathways that were commonly disrupted across the patients, but it also enabled the detection of a second disrupted gene outside the causative locus that is thought to potentially explain the increased severity of the corresponding patient's phenotype. Overall, I highlight the utility of RNA-seq in a clinical setting and add evidence for its incorporation into the diagnostic routine.

However, as a relatively expensive and specialised form of analysis there remain significant barriers to the widespread use of transcriptomics in clinical diagnostics, and this raises questions as to how to select cases most likely to benefit from the approach. With this in mind, it is noteworthy that suspected mitochondrial disorders represent an attractive class of disease (24, 25). With variants in over 340 nuclear genes known to cause mitochondrial disease, the transcriptome-wide nature of RNA-seq has considerable value. However, for other diseases such as tuberous sclerosis where mutations in *TSC1* and *TSC2* account for 95% of clinically suspected cases, it is likely that the targeted generation of RNA-seq data could be more cost-effective and robust (126–129). Furthermore, the genetic heterogeneity of mitochondrial diseases also has advantages for the generation of control RNA-seq data. As demonstrated in this thesis and within other studies, each patient can be analysed with the remainder of the cohort acting as controls reducing the overall cost of sample collection, maintenance of cell lines and sequencing. However, for a disorder for whereby only a handful of genes explain most cases, this would not be advised since common gene mutations/changes would make the identification of outlier expression or splicing events highly challenging.

The selection of control samples is also of broader importance for the outlier detection approaches used in this chapter. Within diagnostics, where pathogenic variants are often unique to a family or patient, the majority of tools developed, including my own, have adopted a 1-vs-all experimental structure (33, 112, 130). However, there remains no consensus on the selection of appropriate control samples for any given set of patients. In certain studies, controls have been matched to the patients on the basis of sex, age and sequencing parameters (23). Others have ignored the demographics of the individuals and adopted a strategy of using the remaining patients in a cohort as controls (24). With the emergence of publicly available datasets containing tens of thousands of RNA-seq samples, it is of particular interest how the number and types of control samples will impact on the detection of aberrant events. Previous work performed with 504 GTEx samples suggests that the number of controls is partially able to overcome the difference in demographic and sequencing protocol between samples, however this remains unclear at larger N numbers (112). The selection of control samples will not only affect the outcome of a diagnostic analysis, but also the practical implementation of RNA-seq in the clinic.

For instance, the use of publicly available data has far fewer associated costs, enables patient samples to be processed flexibly and would permit a standardised protocol across diagnostic laboratories. In contrast, more stringent matching of control and patient samples as well as the use of in-house control samples increases patient recruitment time and costs. Overall, more studies are needed to benchmark the selection of control samples on the detection of aberrant events and this could impact on the results of analyses such as that presented in chapter.

The availability of public control RNA-seq sample sets is also closely related to the choice of patient tissue for RNA-seq analyses. For most diseases the tissue analysed will not be that in which the disease manifests with the notable exception of skeletal muscle disorders. The most used proxy tissues to date have been blood and fibroblasts with concerns raised about the use of both tissue types. One concern is that the proxy tissue will not adequately represent and thus, miss the pathogenic event observed in the disease tissue. Resources such as GTEx and MAJIQ-CAT allow for the assessment of the expression and common splicing of disease genes in clinically accessible tissues (30, 51). Using this approach, recent studies have suggested that the majority of OMIM-genes are expressed within fibroblasts, supporting the choice of this tissue for the analyses presented in this chapter (25). Nonetheless, it has been shown that whole blood RNA-seq can be used to diagnose various diseases including those with a neurological phenotype with a 10% diagnostic yield (27).

As demonstrated in this thesis, prediction of the functional consequence of aberrant splicing events detected through RNA-seq can be used to help interpret their pathogenicity. This process has also been shown by other studies to be useful for informing the development of splice-modulating treatments (44). To date, approaches that interpret the functional consequence of splicing events from short-read RNA-seq require comparison to a reference transcript. The selection of this reference remains challenging, with the vast majority of protein-coding genes presenting with multiple isoforms according to Ensembl annotation (12). This has triggered the development of the MANE-select set of transcripts, which I used to interpret the aberrant splicing events in this thesis. The MANE-select project has the goal of simplifying each gene to a single, functional transcript in most cases. However, this is unlikely to reflect the underlying biology and thus, may lead to the incorrect prediction of the consequence of aberrant splicing events. Furthermore, short-read RNA-seq and more recently, long-read RNA-seq studies elude to the plethora of transcripts left out of annotation databases, which complicates interpretation further (19, 131, 132). Whilst the function and redundancy of these newly discovered transcripts remains contested, it is becoming increasingly likely that the reduction of all genes to a single transcript represents an oversimplification (61). In the future, it is likely that the application of long-read RNA-seq in a diagnostic setting, which can capture full-length transcript structure accurately, will improve the interpretation of the consequences of aberrant splicing events.

Beyond diagnosis, RNA-sequencing also has the ability to improve the understanding of disease pathogenesis through the detection of disruptions to global pathways and genetic contributors to disease outside of the primary locus. For example, I demonstrate that the application of RNA-seq to a cohort of patients with *ATG7* mutations enables the identification of pathways that would be expected to be perturbed such as those related to golgi-function. Furthermore, I identify pathways which have not previously been linked to the *ATG7* function but, given the neurodevelopmental phenotype of these patients, may be important (for example, cell-cycle perturbations). One application of this downstream pathway information would be to help prioritise the causative genes in patients with similar clinical phenotypes. In addition, this approach may be useful for identifying the pathways that could be modulated to ameliorate disease-related processes. However, the global disruptions in RNA expression, which are secondary to the pathogenic event, may also complicate analyses. For instance, the transcriptional changes that are attributed to these pathway disruptions may, in certain cases, become noise that obscures the identification of a causative RNA-level event.

Overall, in this chapter, I add to the accumulating evidence that RNA-seq has utility as a diagnostic tool as well as demonstrate that RNA-seq can improve disease understanding, namely through the characterisation of splicing events, detection of perturbed pathways and potential genetic modifiers outside of the causative locus.

Chapter 5

Conclusions and future directions

5.1 Summary of the thesis contributions

The overarching goal of this thesis is to explore the use of transcriptomics to improve the diagnostic rate of Mendelian disorders. I contribute to this goal by developing methods to improve variant interpretation as well as conducting analyses that support the value of using RNA-sequencing (RNA-seq) for diagnostics and beyond. In chapter 1, I demonstrate that existing gene annotation remains incomplete, which hinders variant interpretation and likely limits diagnostic yield as a consequence. To address this, I leverage publicly available RNA-seq data across 46 human tissues to detect novel exons and improve annotation for the majority of known Mendelian disease genes. This novel annotation is publicly released through a web interface, *vizER*, and the method for its generation as a Bioconductor package, *ODER* (131, 133). In chapter 2, I develop an aberrant sequencing detection method, *dasper*, which uses junction and coverage information from patient-derived RNA-seq data to detect aberrant splicing events. *dasper* is also publicly released via the Bioconductor project and has been designed with diagnostics in mind (112, 134). It applies a 1-vs-all experimental framework, produces a ranked list of aberrant splicing events and allows for the flexible visualisation of aberrant splicing events in the form of sashimi plots. In the final chapter, I apply RNA-seq to patient-derived samples in order to diagnose a set of genetically unsolved cases suspected of having mitochondrial disease due to nuclear mutations. RNA-seq achieved a genetic diagnosis in 1/32 (3.13%) patients and discovered candidate genes for a remaining third, supporting the increasing evidence of the utility of RNA-seq as a diagnostic tool. Each of the above projects has its own discussion, therefore this chapter aims to summarise the major insights obtained from this thesis, highlight the limitations of the work and discuss the future directions of the field of RNA-seq for diagnostics.

One of the key findings from this thesis is the detection of widespread, unannotated transcription across all human tissues, with the implication that gene annotation remains incomplete even for well-studied, disease-associated genes. During the timeline of this PhD, other studies employing computational transcript assembly approaches or long-read sequencing have been published, supporting this finding through the discovery of a plethora of unannotated, human transcripts (19, 20, 132). However, one area not addressed by these studies is the disproportionate impact incomplete annotation has on specific tissues and diseases. I tackle this by comparing the abundance of novel transcription across human tissues and genes. I find that the transcriptome of the brain remains the least well understood and that genes associated with neurodegenerative diseases are enriched amongst the set of genes I re-annotate. This suggests that variant interpretation will be more difficult for genes expressed in the brain. As a consequence, I anticipate that improvements to gene annotation, such as the public release of the novel annotation derived within this thesis, will have the greatest impact on the diagnostic yield of patients with neurogenetic disorders.

In chapter 2, I demonstrate that *dasper* detects pathogenic splicing events more accurately than the existing method, *LeafCutterMD* (35). Other similar aberrant splicing detection tools have been released during the course of this PhD, namely *FRASER* and *SPOT* (36, 135). It worth noting that *SPOT* is not directly designed for diagnostics, rather the detection of aberrant splicing events in unaffected individuals. Therefore, it remains unclear how *SPOT* would perform on patients with Mendelian disorders, where there are likely to be global transcriptional changes related to disease pathogenesis that could obscure the signal of the pathogenic event. However, the outlier detection models employed by *FRASER* and *SPOT* are both comparable to *dasper*s therefore, in future work it would be useful to benchmark the performance of *dasper* to these tools. Despite the growing number of aberrant splicing detection tools, *dasper* addresses three areas that, to my knowledge, have not been formally tackled by any other splicing method in the diagnostic RNA-seq space. Firstly, analyses using *dasper* revealed that the incorporation of coverage data alongside junction counts can be used to improve the detection of aberrant splicing events. Coverage information has been employed by differential splicing tools such as *dSpliceType* (136). However, *dasper* is the only tool applied to RNA-seq for diagnostics that demonstrates the utility of coverage within an N-vs-1 experiment, where one would expect the variance of coverage noise to be greater. Secondly, *dasper* includes functionality for visualising aberrant splicing events in the form of sashimi plots, which facilitates the interpretation of the consequence of these events with respect to transcript structure and function. The addition of plotting functionality within *dasper* itself permits a more efficient, automated diagnostics pipeline, streamlining the processes of detection and interpretation. Thirdly, I compare the detection of known pathogenic splicing events through *dasper* using in-house samples versus publicly available RNA-seq data as controls. To my knowledge, this has not yet been investigated by any of the other aberrant splicing detection methods, which have relied on in-house controls generated using similar sequencing protocols to the patient samples. Further exploration of the impact of using publicly available data as controls is likely to be extremely valuable for several reasons. The use of publicly available data can: i) reduce the cost of the diagnostic workflow by avoiding the need to sequence control samples making the analysis more affordable, ii) standardise the diagnostic RNA-seq pipeline between laboratories, which would otherwise be using different control data sets, and iii) avoid the need to batch patient RNA-seq samples and therefore delays to genetic diagnoses.

In chapter 3, the application of RNA-seq for diagnosing a cohort of unsolved cases with suspected mitochondrial disorders supports the accumulating evidence that RNA-seq serves as a promising diagnostic tool. The diagnostic success rate I obtain, 3.15%, is similar to the existing, diagnostic RNA-seq studies on patients with mitochondrial disorders (24, 25). Given the requirement for more standardised pipelines within the diagnostic RNA-seq field, I outline a workflow for the prioritisation of candidate genes using detection of aberrant splicing and expression events from RNA-seq. During the course of this PhD, Yépez and colleagues have released an open-source, modular workflow, *DROP*, which also incorporates ASE analyses and variant calling from RNA-seq (85). I anticipate that *DROP* is likely to be immensely valuable to clinicians and researchers, representing a milestone in the movement towards the standardisation and automation of diagnostic RNA-seq, which will also greatly aid the adoption of RNA-seq within diagnostic laboratories. Notably, the modular design of *DROP* permits the extension of the workflow in future. In particular, as an area of growing interest, this could include the addition of the use of RNA-seq to inform the development of personalised therapies. For instance, as demonstrated in chapter 3, RNA-seq can be used to characterise the consequences of aberrant splicing events, a process that has been shown to be helpful for the design of personalised, ASO therapies (44). Overall, the release of standardised workflows such as *DROP* are a major step towards the adoption of RNA-seq within the diagnostic routine; it is my hope that in the future, such workflows will also incorporate tools that use RNA-seq to benefit patients beyond diagnoses.

5.2 Limitations of RNA-seq for diagnostics

While the analyses conducted within my thesis provide valuable contributions to the field of diagnostic RNA-seq, there are several limitations. These limitations arise from the current incomplete understanding of transcriptome complexity, the rapid evolution of sequencing technologies, the relative paucity of publicly available patient RNA-seq data and differences in transcript expression and/or splicing between human tissues.

One of the key issues raised in this thesis is the requirement for accurate transcript structures as a basis to interpret the impact of a given variant or splicing event. For any given gene, there can be a large number of potential transcript structures reported, which could potentially generate conflicting predictions of pathogenicity (21). Furthermore, since aberrant splicing is called locally using short read RNA-seq data rather than within the context of a whole transcript, I have made a key assumption in order to perform the analysis, namely that there is a major transcript structure of interest and that the aberrant event occurs within that structure. More generally, there is a drive to simplify transcript annotation as exemplified by resources such as MANE-select and methods such as *APPRIS*, which aim to prioritise 1 or 2 principle, functional transcripts per gene (137, 138). Albeit useful for the standardisation of variant interpretation across analyses and annotation databases, these approaches may lead to a substantial oversimplification of transcriptome complexity. As a result, these approaches can lead to erroneous predictions and inaccurate variant interpretation, in particular for variants that lie in non-coding regions with respect to the MANE-select transcript. For example, in the case of the gene *TTN*, the exonic usage and isoform expression of this gene are tightly regulated among different developmental and physiological states. In fact, a subset of *TTN* exons are only expressed during development and never found in adult human skeletal muscle (139). This suggests that at least a subset of genes are likely to have more than a single functional transcript, illustrating the need to account for transcript complexity in order to accurately interpret variants (140). The advent of 3rd generation sequencing technologies such as Oxford Nanopore and Pacific Biosciences, which have the capability to capture full-length transcript structures, will provide a more accurate view of transcriptome complexity (65, 66). Therefore, long-read RNA-seq is likely to have a revolutionary impact on the development of gene annotation databases. In particular, as the sequencing depth remains a major bottleneck for whole transcriptome long-read sequencing, targeted approaches have been employed which are capable of improving the transcript resolution for complex genes such as *TTN* and *CACNA1C* (20, 67). In the context of diagnostic RNA-seq, targeted long-read RNA-seq can improve variant interpretation in two ways. When applied to control samples, targeted long-read RNA-seq can be used to profile transcriptome complexity and improve the accuracy of reference annotation databases. When applied to patient samples, these technologies allow you to directly measure changes to the full-length transcript structure and relative abundance allowing more accurate detection of aberrant events and derivation of their functional consequence. Overall, it is my hope that long-read sequencing represents a new phase for diagnostics, whereby the fundamental unit used for interpretation will shift from the existing gene-centric paradigm or local, cluster-based approaches, to a transcript-oriented view that more accurately represents the underlying biology.

Sequencing technologies have evolved not only in the length of outputted reads but also the resolution and type of their inputted samples. For instance, single-cell RNA-seq is an emerging technology that allows profiling the transcriptome at cellular resolution (141). In addition, there has been an increase in the number of publicly available datasets such as that released by the BrainSEQ consortium, PsychENCODE, the Human Developmental Cell Atlas and SCDevDB that include samples from various stages of human development (68–70, 142). These examples highlight a limitation of the analyses within this thesis, namely that it is applied solely on bulk tissue samples from human adults. In chapter 1, I found that novel transcription was most commonly discovered in the human brain, which is likely attributed to the relative lack of brain-derived data that has entered annotation pipelines to date. Accordingly, one would also expect that single-cell and developmental RNA-seq datasets would be enriched for previously unannotated cell or developmentally specific transcripts. I anticipate that the future application of approaches such as *ODER* and long-read sequencing on these datasets will be valuable for improving our understanding of the transcriptomic changes that account for the differences between cell-types and developmental stages. In fact, recent studies have already started to address this by combining long-read and single cell RNA-seq technologies (143, 144).

Despite the progressively increasing quantity of publicly available RNA-seq data, this has not encompassed data derived from patient samples. Although there are several published studies applying diagnostic RNA-seq on hundreds of samples, to my knowledge, none of these release their RNA-seq data publicly (23–25, 27). Predominantly, this is due to static consent frameworks that restrict the use of patient samples to research projects outlined at the stage of sample retrieval (145). Following from this, a key limitation of the development of *dasper* in chapter 2 is the reliance on a small number of patient samples with pathogenic splicing variants. Importantly, this limitation is not only related to the total number of patient samples, but also the type of pathogenic splicing variant carried by each patient. For the purpose of developing an aberrant splicing detection method, samples from patients carrying deep intronic mutations would be most valuable, however this class of variant itself forms only a small proportion of pathogenic variant databases such as ClinVar (146). The small N number of patient samples limits the generalisability of *dasper*, as the detection of aberrant splicing events is influenced by the type of pathogenic splicing mutation, the disorder of the patient and the demographic of the samples analysed. Without a larger cohort of positive controls, it is difficult to extrapolate the performance of *dasper* across these variables. This problem of benchmarking *dasper* is compounded, as the small N number also introduces testing circularity; the 16 patient samples which inform *dasper*'s development are the same 16 which are used to validate its performance. In addition, the design of the computational method within *dasper* relies on the assumption of both an up- and down-regulated junction to be present in each pathogenic splicing event. Although this holds true for the 16 positive controls on which *dasper* is tested, given the limited sample number it is unclear whether this assumption will remain the case for all pathogenic splicing events. Notably, this could be a particular issue for intron retention events, which are only represented by a single patient in the analysed cohort, wherein one might not expect the appearance of an up-regulated junction. It is likely that the quantity of publicly available patient samples will increase in the future for several reasons. First, new dynamic consent policies employed by large organisations such as Genomics England in the UK allow extensions to the research scope that a patient sample can be used for (145). Secondly, there is an increasing adoption of cloud-based, computational infrastructures such as Terra, which enable access to vast compute resources and data to be processed without being downloaded (147, 148). Finally, many aberrant splicing detection tools do not require read-level information, which carries the risk of an individual being identified from their genetic variants. Instead, they can leverage non-identifiable junction count matrices and BigWig files as input; these data formats can be shared more readily. This is exemplified by the recount projects which standardise processing and publicly release RNA-seq data from hundreds of thousands of samples (52, 63). Together, it is my hope that these factors will lead to easier access to a larger number of patient-derived RNA-seq data in the future, which will aid the development of better quality bioinformatics tools. In the case of *dasper*, this will allow the performance and assumptions of the method to be assessed across a larger cohort of patients and specifically, those that carry deep intronic splicing variants of interest.

While the development of the tools such as *dasper* and workflows such as *DROP* are valuable within research studies, there are still bottlenecks that prevent these approaches from being widely adopted in clinical settings (85, 112). One major hurdle is the lack of clear, community-accepted criteria for evaluating RNA-seq evidence in a diagnostic context. Without such guidelines, it remains the responsibility of individual groups and clinicians to assess the evidence resulting from the applied methods and workflows. This increases the time taken for interpretation and leads to a set of diverse, subjective criteria or thresholds being employed for determining the pathogenicity of RNA events that varies between studies and research groups. Although this is surmountable within individual research studies, it represents a much greater bottleneck for diagnostic laboratories. It is likely a set of extended clinical guidelines, similar to the update of the ACMG guidelines driven by the adoption of WES as a standard diagnostic approach, will be required for RNA-seq (149). The development of these criteria should involve a conversation between the community and consider the effect sizes and statistical cut-offs that would define a pathological or benign phenotype from RNA-seq data. For example, a threshold of observed population frequency could be established for junctions, above which the splicing event represented by that junction should be considered benign. It is worth noting that such cut-offs would need to be more complex than those currently used for genetic data, for example by accounting for tissue, cell-type and developmental differences. This is likely to require a better understanding of the landscape of normal junction frequencies than currently available. Ultimately, this would greatly aid the adoption of RNA-seq within a clinical setting by standardising the criteria for assigning the pathogenicity of variants using RNA-seq based evidence.

Another core limitation of diagnostic RNA-seq is that often, the sampled tissue from a patient is not a disease-relevant tissue. This phenomena is commonly termed the proxy tissue problem and can arise for multiple reasons (30). The disease tissue may be physically inaccessible without harming the patient, for example in the case of neurological disorders. For developmental disorders, certain disease changes may only be observable at a specific developmental stage, which cannot be sampled in adult patients. Finally, in certain disorders, the transcriptional signature of disease may be constrained to a specific cell-type, and therefore, masked in bulk tissue samples. In all cases, the proxy tissue problem leads to major challenges, whereby the disease-associated transcriptional changes may not be adequately captured in the proxy tissue. Due to the variability in gene expression across human tissues, it is unclear whether the disease-associated transcript will be expressed (and therefore captured) within a proxy tissue. Even if expressed, the splicing and transcript-level changes related to disease state may not be represented in a proxy tissue. Furthermore, for practical reasons, often a cell line is used as the proxy tissue. However, the *in vitro* culturing and maintenance of cell lines induces transcriptional changes that may not accurately reflect the endogenous, disease state. It is important to note that proxy tissue problem will impact different diseases to a varying degree. For example, skeletal muscle disorders such as Duchenne muscular dystrophy are an ideal disease for application of diagnostic RNA-seq as muscle biopsies are already routinely performed as part of the diagnostic pathway. As a consequence, studies analysing patients with skeletal disorders, to my knowledge, have achieved the highest diagnostic yield obtained by application of RNA-seq to date (35%) (23). In chapter 3, I apply RNA-seq to diagnose patients with suspected mitochondrial disorders using data derived from fibroblast cell lines. Disrupted mitochondrial function is likely to have impact on transcription ubiquitously across tissues, however the degree of this impact is likely to vary due to tissue differences in expression and splicing. In addition, transcriptional artefacts are likely to be induced during the culturing process of fibroblasts. Together, it is likely that the proxy tissue problem still limits the diagnostic yield of RNA-seq applied to mitochondrial disease patients. It is foreseeable that the proxy tissue problem can be overcome through technological developments, such as the differentiation or transdifferentiation of cell lines into the disease-relevant tissues. In fact, research studies have employed such techniques successfully in the diagnostic context, for example through the differentiation of IPSCs to retinal organoids or transdifferentiation of fibroblasts to myoblasts (28, 150). These examples suggest that the conversion of proxy tissues to a disease relevant tissue can reproduce disease-associated transcriptional changes that would otherwise have been missed. Although these technologies represent solution for disorders where the disease tissue remains inaccessible, they are not without drawbacks. Primarily, they incur an increased resource and expertise burden, which prevent them being from being widely adopted in a clinical setting. Additionally, the differentiation process will generate genetic and transcriptomic artefacts which may obscure disease associated events. Overall, the proxy tissue problem remains a major challenge, which limits diagnostic yield obtainable through RNA-seq, especially for disorders where the disease tissue is inaccessible. Although technological advances in differentiating cell lines represents a viable solution in principle, their incorporation into the diagnostic routine is not practically feasible until their costs can be greatly reduced.

Given the considerable resources required to generate patient-derived RNA-seq data this also raises questions about whether this form of data provides additional insights beyond diagnosis for the individual patients or patient cohorts.

5.3 RNA-sequencing beyond diagnostics

The clinical value of an RNA-seq experiment lies not only in establishing a genetic diagnosis for patients. RNA-seq data can provide additional information about disease pathogenesis, which could in turn lead to more accurate prognostic information and the development of personalised therapies. More specifically, as demonstrated in chapter 3, RNA-seq data can be used to characterise the pathogenesis of Mendelian disorders by detecting transcriptomic changes: i) at the causative locus, and ii) more broadly, outside of the causative locus. Focusing on the former, characterisation of aberrant splicing events, can inform the design of splice-modulating therapies such as ASOs. For example, I demonstrate that junction and coverage information from RNA-seq data can be used to classify aberrant splicing events by their consequence on the transcript structure. In previous studies, a similar approach has been successfully employed to design ASOs targeting a pathogenic splicing event in a patient with Batten's disease (44). Together with the recent development of other successful splice-modulating therapies, such as the FDA-approved oligonucleotide, nusinersen for spinal muscular atrophy, the area of RNA modulating compounds is of growing interest amongst drug discovery companies (151, 152).

Given that transcriptomic data provides a broad functional readout of cellular activity, it potentially contains information about downstream pathways that are perturbed in disease and genetic modifiers contributing to severity. In the field of cancer research, the value of understanding the genetic profile and disrupted transcriptomic pathways downstream of tumour genesis are already well-recognised. For example, pathway and variant level information has been combined with clinical symptoms to inform the prognosis of patients with breast cancer (153, 154). In addition, the molecular profiling of tumour cells has become a fundamental component of the development of precision cancer treatments tailored to the somatic mosaicism within individual patients (155). There are reasons to believe that these approaches should also be applicable to field of rare diseases. Pathway activation has been shown to correlate with phenotype severity in Mendelian disorders. For instance, in cystic fibrosis patients the expression of genes within the type I interferon response pathway have been demonstrated to be a determinant of whether a patient will have mild or severe lung phenotype (93). Similarly, studies using mutational burden analyses to investigate sickle cell anemia have found that variants in *CLCN6* or *OGHDL* were enriched in patients that had a longer survival duration (92). With 5% of Mendelian patients observed to have variation at multiple loci contributing to major phenotypic features, the importance of such genetic modifiers is likely to be underappreciated (91).

Overall, I anticipate that the full clinical value of RNA-seq will not only be for genetically diagnosing a patient, but also to improve understanding of their disease pathogenesis and aid the design of personalised, RNA-modulating therapies.

5.4 Concluding remarks

The evidence from this thesis and other studies released during the course of my PhD have established RNA-seq as a valuable tool for improving the genetic diagnostic rate of Mendelian disease patients. I expect that, over the next decade, the emerging diagnostic RNA-seq workflows will mature and be supplemented with the release of a standardised, diagnostic criteria for evaluating RNA-seq evidence. This will enable the adoption of RNA-seq as part of the staple diagnostic routine. In parallel, with the growth in the field of personalised medicine, I envision that the clinical application of RNA-seq will evolve beyond diagnostics, to become a framework for improving disease understanding as well as a platform for the discovery of RNA-modulating therapies.

References

[1] INSERM. *Prevalence and incidence of rare diseases : Bibilographic data.* 2015.

[2] Kym M. Boycott, Ana Rath, Jessica X. Chong, Taila Hartley, Fowzan S. Alkuraya, Gareth Baynam, Anthony J. Brookes, Michael Brudno, Angel Carracedo, Johan T. den Dunnen, Stephanie O.M. Dyke, Xavier Estivill, Jack Goldblatt, Catherine Gonthier, Stephen C. Groft, Ivo Gut, Ada Hamosh, Philip Hieter, Sophie Höhn, Matthew E. Hurles, Petra Kaufmann, Bartha M. Knoppers, Jeffrey P. Krischer, Milan Macek, Gert Matthijs, Annie Olry, Samantha Parker, Justin Paschall, Anthony A. Philippakis, Heidi L. Rehm, Peter N. Robinson, Pak Chung Sham, Rumen Stefanov, Domenica Taruscio, Divya Unni, Megan R. Vanstone, Feng Zhang, Han Brunner, Michael J. Bamshad, and Hanns Lochmüller. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *American Journal of Human Genetics*, 100(5):695–705, 2017. ISSN 15376605. doi: 10.1016/j.ajhg.2017.04.003.

[3] Laure Frésard and Stephen B Montgomery. Diagnosing rare diseases after the exome, 2018. ISSN 23732873.

[4] Hane Lee, Alden Y. Huang, Lee kai Wang, Amanda J. Yoon, Genecee Renteria, Ascia Eskin, Rebecca H. Signer, Naghmeh Dorrani, Shirley Nieves-Rodriguez, Jijun Wan, Emilie D. Douine, Jeremy D. Woods, Esteban C. Dell'Angelica, Brent L. Fogel, Martin G. Martin, Manish J. Butte, Neil H. Parker, Richard T. Wang, Perry B. Shieh, Derek A. Wong, Natalie Gallant, Kathryn E. Singh, Y. Jane Tavyev Asher, Janet S. Sinsheimer, Deborah Krakow, Sandra K. Loo, Patrick Allard, Jeanette C. Papp, Christina G.S. Palmer, Julian A. Martinez-Agosto, and Stanley F. Nelson. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genetics in Medicine*, 22(3):490–499, 2020. ISSN 15300366. doi: 10.1038/s41436-019-0672-1. URL http://dx.doi.org/10.1038/s41436-019-0672-1.

[5] 100,000 Genomes Project Pilot Investigators, Damian Smedley, Katherine R Smith, Antonio Martin, Ellen A Thomas, Ellen M McDonagh, Valentina Cipriani, Jamie M Ellingford, Gavin Arno, Arianna Tucci, Jana Vandrovcova, Georgia Chan, Hywel J Williams, Thiloka Ratnaike, Wei Wei, Kathleen Stirrups, Kristina Ibanez, Loukas Moutsianas, Matthias Wielscher, Anna Need, Michael R Barnes, Letizia Vestito, James Buchanan, Sarah Wordsworth, Sofie Ashford, Karola Rehmström, Emily Li, Gavin Fuller, Philip Twiss, Olivera Spasic-Boskovic, Sally Halsall, R Andres Floto, Kenneth Poole, Annette Wagner, Sarju G Mehta, Mark Gurnell, Nigel Burrows, Roger James, Christopher Penkett, Eleanor Dewhurst, Stefan Gräf, Rutendo Mapeta, Mary Kasanicki, Andrea Haworth, Helen Savage, Melanie Babcock, Martin G Reese, Mark Bale, Emma Baple, Christopher Boustred, Helen Brittain, Anna de Burca, Marta Bleda, Andrew Devereau, Dina Halai, Eik Haraldsdottir, Zerin Hyder, Dalia Kasperaviciute, Christine Patch, Dimitris Polychronopoulos, Angela Matchan, Razvan Sultana, Mina Ryten, Ana L T Tavares, Carolyn Tregidgo, Clare Turnbull, Matthew Welland, Suzanne Wood, Catherine Snow, Eleanor Williams, Sarah Leigh, Rebecca E Foulger, Louise C Daugherty, Olivia Niblock, Ivone U S Leong, Caroline F Wright, Jim Davies, Charles Crichton, James Welch, Kerrie Woods, Lara Abulhoul, Paul Aurora, Detlef Bockenhauer, Alexander Broomfield, Maureen A Cleary, Tanya Lam, Mehul Dattani, Emma Footitt, Vijeya Ganesan, Stephanie Grunewald, Sandrine Compeyrot-Lacassagne, Francesco Muntoni, Clarissa Pilkington, Rosaline Quinlivan, Nikhil Thapar, Colin Wallis, Lucy R Wedderburn, Austen Worth, Teofila Bueser, Cecilia Compton, Charu Deshpande, Hiva Fassihi, Eshika Haque, Louise Izatt, Dragana

Josifova, Shehla Mohammed, Leema Robert, Sarah Rose, Deborah Ruddy, Robert Sarkany, Genevieve Say, Adam C Shaw, Agata Wolejko, Bishoy Habib, Gavin Burns, Sarah Hunter, Russell J Grocock, Sean J Humphray, Peter N Robinson, Melissa Haendel, Michael A Simpson, Siddharth Banka, Jill Clayton-Smith, Sofia Douzgou, Georgina Hall, Huw B Thomas, Raymond T O'Keefe, Michel Michaelides, Anthony T Moore, Sam Malka, Nikolas Pontikos, Andrew C Browning, Volker Straub, Gráinne S Gorman, Rita Horvath, Richard Quinton, Andrew M Schaefer, Patrick Yu-Wai-Man, Doug M Turnbull, Robert McFarland, Robert W Taylor, Emer O'Connor, Janice Yip, Katrina Newland, Huw R Morris, James Polke, Nicholas W Wood, Carolyn Campbell, Carme Camps, Kate Gibson, Nils Koelling, Tracy Lester, Andrea H Németh, Claire Palles, Smita Patel, Noemi B A Roy, Arjune Sen, John Taylor, Pilar Cacheiro, Julius O Jacobsen, Eleanor G Seaby, Val Davison, Lyn Chitty, Angela Douglas, Kikkeri Naresh, Dom McMullan, Sian Ellard, I Karen Temple, Andrew D Mumford, Gill Wilson, Phil Beales, Maria Bitner-Glindzicz, Graeme Black, John R Bradley, Paul Brennan, John Burn, Patrick F Chinnery, Perry Elliott, Frances Flinter, Henry Houlden, Melita Irving, William Newman, Shamima Rahman, John A Sayer, Jenny C Taylor, Andrew R Webster, Andrew O M Wilkie, Willem H Ouwehand, F Lucy Raymond, John Chisholm, Sue Hill, David Bentley, Richard H Scott, Tom Fowler, Augusto Rendon, and Mark Caulfield. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *The New England journal of medicine*, 385(20):1868–1880, 2021. ISSN 1533-4406. doi: 10.1056/NEJMoa2035790. URL http://www.ncbi.nlm.nih.gov/pubmed/34758253.

[6] D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–76, apr 2014. ISSN 1476-4687. doi: 10.1038/nature13127. URL http://dx.doi.org/10.1038/nature13127http://www.nature.com/articles/nature13127http://www.ncbi.nlm.nih.gov/pubmed/24759409http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4180223.

[7] Sierra S Nishizaki and Alan P Boyle. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms, 2017. ISSN 13624555. URL http://dx.doi.org/10.1016/j.tig.2016.10.008.

[8] Rita Vaz-Drago, Noélia Custódio, and Maria Carmo-Fonseca. Deep intronic mutations and human disease. *Human Genetics*, 136(9):1093–1111, 2017. ISSN 14321203. doi: 10.1007/s00439-017-1809-4.

[9] Ying Hu, Payam Mohassel, Sandra Donkervoort, Pomi Yun, Véronique Bolduc, Daniel Ezzo, Jahannaz Dastgir, Jamie L. Marshall, Monkol Lek, Daniel G. MacArthur, A. Reghan Foley, and Carsten G. Bönnemann. Identification of a Novel Deep Intronic Mutation in CAPN3 Presenting a Promising Target for Therapeutic Splice Modulation. *Journal of Neuromuscular Diseases*, 53 (3):1–9, aug 2019. ISSN 22143599. doi: 10.3233/JND-190414. URL http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JAD-160280https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JND-190414.

[10] Yaping Yang, Donna M. Muzny, Fan Xia, Zhiyv Niu, Richard Person, Yan Ding, Patricia Ward, Alicia Braxton, Min Wang, Christian Buhay, Narayanan Veeraraghavan, Alicia Hawes, Theodore Chiang, Magalie Leduc, Joke Beuten, Jing Zhang, Weimin He, Jennifer Scull, Alecia Willis, Megan Landsverk, William J. Craigen, Mir Reza Bekheirnia, Asbjorg Stray-Pedersen, Pengfei Liu, Shu Wen, Wendy Alcaraz, Hong Cui, Magdalena Walkiewicz, Jeffrey Reid, Matthew Bainbridge, Ankita Patel, Eric Boerwinkle, Arthur L. Beaudet, James R. Lupski, Sharon E. Plon, Richard A. Gibbs, and Christine M. Eng. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA*, 312 (18):1870, nov 2014. ISSN 0098-7484. doi: 10.1001/jama.2014.14601. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2014.14601.

[11] Jenny C. Taylor, Hilary C. Martin, Stefano Lise, John Broxholme, Jean Baptiste Cazier, Andy Rimmer, Alexander Kanapin, Gerton Lunter, Simon Fiddy, Chris Allan, A. Radu Aricescu, Moustafa Attar, Christian Babbs, Jennifer Becq, David Beeson, Celeste Bento, Patricia Bignell, Edward Blair, Veronica J. Buckle, Katherine Bull, Ondrej Cais, Holger Cario, Helen Chapel, Richard R. Copley, Richard Cornall, Jude Craft, Karin Dahan, Emma E. Davenport, Calliope Dendrou, Olivier Devuyst, Aimée L. Fenwick, Jonathan Flint, Lars Fugger, Rodney D. Gilbert, Anne Goriely, Angie Green, Ingo H. Greger, Russell Grocock, Anja V. Gruszczyk, Robert Hastings, Edouard Hatton, Doug Higgs, Adrian Hill, Chris Holmes, Malcolm Howard, Linda Hughes, Peter Humburg, David Johnson, Fredrik Karpe, Zoya Kingsbury, Usha Kini, Julian C. Knight, Jonathan Krohn, Sarah Lamble, Craig Langman, Lorne Lonie, Joshua Luck, Davis McCarthy, Simon J. McGowan, Mary Frances McMullin, Kerry A. Miller, Lisa Murray, Andrea H. Németh, M. Andrew Nesbit, David Nutt, Elizabeth Ormondroyd, Annette Bang Oturai, Alistair Pagnamenta, Smita Y. Patel, Melanie Percy, Nayia Petousi, Paolo Piazza, Sian E. Piret, Guadalupe Polanco-Echeverry, Niko Popitsch, Fiona Powrie, Chris Pugh, Lynn Quek, Peter A. Robbins, Kathryn Robson, Alexandra Russo, Natasha Sahgal, Pauline A. Van Schouwenburg, Anna Schuh, Earl Silverman, Alison Simmons, Per Soelberg Sorensen, Elizabeth Sweeney, John Taylor, Rajesh V. Thakker, Ian Tomlinson, Amy Trebes, Stephen R.F. Twigg, Holm H. Uhlig, Paresh Vyas, Tim Vyse, Steven A. Wall, Hugh Watkins, Michael P. Whyte, Lorna Witty, Ben Wright, Chris Yau, David Buck, Sean Humphray, Peter J. Ratcliffe, John I. Bell, Andrew O.M. Wilkie, David Bentley, Peter Donnelly, and Gilean McVean. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, 47(7):717–726, 2015. ISSN 15461718. doi: 10.1038/ng.3304. URL http://dx.doi.org/10.1038/ng.3304.

[12] Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018. ISSN 13624962. doi: 10.1093/nar/gkx1098.

[13] J Harrow, A Frankish, J M Gonzalez, E Tapanari, M Diekhans, F Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22 (9):1760–1774, sep 2012. ISSN 1088-9051. doi: 10.1101/gr.135350.111. URL https://doi.org/10.1101/gr.135350.111http://genome.cshlp.org/cgi/doi/10.1101/gr.135350.111.

[14] Danielle Thierry-Mieg and Jean Thierry-Mieg. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome biology*, 2006. ISSN 1465-6914. doi: 10.1186/gb-2006-7-s1-s12.

[15] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(SUPPL. 1):61–65, 2007. ISSN 03051048. doi: 10.1093/nar/gkl842.

[16] Geng Chen, Charles Wang, Leming Shi, Xiongfei Qu, Jiwei Chen, Jianmin Yang, Caiping Shi, Long Chen, Peiying Zhou, Baitang Ning, Weida Tong, and Tieliu Shi. Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. *RNA (New York, N.Y.)*, 19(4):479–89, 2013. ISSN 1469-9001. doi: 10.1261/rna.037473.112. URL http://rnajournal.cshlp.org/content/19/4/479.long.

[17] Adam Frankish, Barbara Uszczynska, Graham R S Ritchie, Jose M Gonzalez, Dmitri Pervouchine, Robert Petryszak, Jonathan M Mudge, Nuno Fonseca, Alvis Brazma, Roderic Guigo, and Jennifer Harrow. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*, 16(8):1–11, 2015. ISSN 14712164. doi: 10.1186/1471-2164-16-S8-S2.

[18] Andrew E. Jaffe, Jooheon Shin, Leonardo Collado-Torres, Jeffrey T. Leek, Ran Tao, Chao Li, Yuan Gao, Yankai Jia, Brady J. Maher, Thomas M. Hyde, Joel E. Kleinman, and Daniel R. Weinberger. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Neuroscience*, 18(1):154–161, 2015. ISSN 15461726. doi: 10.1038/nn.3898. URL http://dx.doi.org/10.1038/nn.3898.

[19] Mihaela Pertea, Alaina Shumate, Geo Pertea, Ales Varabyou, Yu-Chi Chang, Anil K Madugundu, Akhilesh Pandey, and Steven Salzberg. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology*, page 332825, 2018. ISSN 1474-760X. doi: 10.1101/332825. URL https://www.biorxiv.org/content/early/2018/05/28/332825?%3Fcollection=.

[20] Michael B. Clark, Tomasz Wrzesinski, Aintzane B. Garcia, Nicola A.L. Hall, Joel E. Kleinman, Thomas Hyde, Daniel R. Weinberger, Paul J. Harrison, Wilfried Haerty, and Elizabeth M. Tunbridge. Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain. *Molecular Psychiatry*, 25(1):37–47, 2020. ISSN 14765578. doi: 10.1038/s41380-019-0583-1. URL http://dx.doi.org/10.1038/s41380-019-0583-1.

[21] Davis J. McCarthy, Peter Humburg, Alexander Kanapin, Manuel A. Rivas, Kyle Gaulton, Jean Baptiste Cazier, and Peter Donnelly. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3), 2014. ISSN 1756994X. doi: 10.1186/gm543.

[22] Samya Chakravorty and Madhuri Hegde. Clinical Utility of Transcriptome Sequencing: Toward a Better Diagnosis for Mendelian Disorders. *Clinical Chemistry*, 000: clinchem.2017.276980, 2017. ISSN 0009-9147. doi: 10.1373/clinchem.2017.276980. URL http://www.clinchem.org/lookup/doi/10.1373/clinchem.2017.276980.

[23] Beryl B. Cummings, Jamie L. Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A. Reghan Foley, Veronique Bolduc, Leigh B. Waddell, Sarah A. Sandaradura, Gina L. O'Grady, Elicia Estrella, Hemakumar M. Reddy, Fengmei Zhao, Ben Weisburd, Konrad J. Karczewski, Anne H. O'Donnell-Luria, Daniel Birnbaum, Anna Sarkozy, Ying Hu, Hernan Gonorazky, Kristl Claeys, Himanshu Joshi, Adam Bournazos, Emily C. Oates, Roula Ghaoui, Mark R. Davis, Nigel G. Laing, Ana Topf, Peter B. Kang, Alan H. Beggs, Kathryn N. North, Volker Straub, James J. Dowling, Francesco Muntoni, Nigel F. Clarke, Sandra T. Cooper, Carsten G. Bönnemann, and Daniel G. MacArthur. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine*, 9(386):eaal5209, 2017. ISSN 1946-6234. doi: 10.1126/scitranslmed.aal5209. URL http://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.aal5209.

[24] Laura S. Kremer, Daniel M. Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B. Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, Eliška Koňaříková, Birgit Repp, Gabi Kastenmüller, Jerzy Adamski, Peter Lichtner, Christoph Leonhardt, Benoit Funalot, Alice Donati, Valeria Tiranti, Anne Lombes, Claude Jardel, Dieter Gläser, Robert W. Taylor, Daniele Ghezzi, Johannes A. Mayr, Agnes Rötig, Peter Freisinger, Felix Distelmaier, Tim M. Strom, Thomas Meitinger, Julien Gagneur,

and Holger Prokisch. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications*, 8:15824, 2017. ISSN 2041-1723. doi: 10.1038/ncomms15824. URL http://www.nature.com/doifinder/10.1038/ncomms15824.

[25] Vicente A Yépez, Mirjana Gusic, Robert Kopajtich, Christian Mertes, H Nicholas, Charlotte L Alston, Rui Ban, Skadi Beblo, Riccardo Berutti, Holger Blessing, Elżbieta Ciara, Felix Distelmaier, Peter Freisinger, Johannes Häberle, Susan J Hayflick, Costanza Lamperti, Dominic Lenz, Christine C Makowski, Signe Mosegaard, Michaela F Müller, Gerard Muñoz-pujol, Agnieszka Nadel, Akira Ohtake, Elena Procopio, Thomas Schwarzmayr, Joél Smet, Christian Staufner, L Sarah, Tim M Strom, Caterina Terrile, Frederic Tort, Rudy Van Coster, Matias Wagner, Manting Xu, Fang Fang, Daniele Ghezzi, and A Johannes. Clinical implementation of RNA sequencing for Mendelian disease diagnostics Affiliations. pages 1–52, 2021.

[26] Guillermo Marco-Puche, Sergio Lois, Javier Benítez, and Juan Carlos Trivino. RNA-Seq Perspectives to Improve Clinical Diagnosis. *Frontiers in Genetics*, 10(November):1–7, 2019. ISSN 16648021. doi: 10.3389/fgene.2019.01152.

[27] Laure Frésard, Craig Smail, Nicole M. Ferraro, Nicole A. Teran, Xin Li, Kevin S. Smith, Devon Bonner, Kristin D. Kernohan, Shruti Marwaha, Zachary Zappala, Brunilda Balliu, Joe R. Davis, Boxiang Liu, Cameron J. Prybol, Jennefer N. Kohler, Diane B. Zastrow, Chloe M. Reuter, Dianna G. Fisk, Megan E. Grove, Jean M. Davidson, Taila Hartley, Ruchi Joshi, Benjamin J. Strober, Sowmithri Utiramerur, David R. Adams, Aaron Aday, Mercedes E. Alejandro, Patrick Allard, Euan A. Ashley, Mahshid S. Azamian, Carlos A. Bacino, Eva Baker, Ashok Balasubramanyam, Hayk Barseghyan, Gabriel F. Batzli, Alan H. Beggs, Babak Behnam, Hugo J. Bellen, Jonathan A. Bernstein, Gerard T. Berry, Anna Bican, David P. Bick, Camille L. Birch, Devon Bonner, Braden E. Boone, Bret L. Bostwick, Lauren C. Briere, Elly Brokamp, Donna M. Brown, Matthew Brush, Elizabeth A. Burke, Lindsay C. Burrage, Manish J. Butte, Shan Chen, Gary D. Clark, Terra R. Coakley, Joy D. Cogan, Heather A. Colley, Cynthia M. Cooper, Heidi Cope, William J. Craigen, Precilla D'Souza, Mariska Davids, Jean M. Davidson, Jyoti G. Dayal, Esteban C. Dell'Angelica, Shweta U. Dhar, Katrina M. Dipple, Laurel A. Donnell-Fink, Naghmeh Dorrani, Daniel C. Dorset, Emilie D. Douine, David D. Draper, Annika M. Dries, Laura Duncan, David J. Eckstein, Lisa T. Emrick, Christine M. Eng, Gregory M. Enns, Ascia Eskin, Cecilia Esteves, Tyra Estwick, Liliana Fernandez, Carlos Ferreira, Elizabeth L. Fieg, Paul G. Fisher, Brent L. Fogel, Noah D. Friedman, William A. Gahl, Emily Glanton, Rena A. Godfrey, Alica M. Goldman, David B. Goldstein, Sarah E. Gould, Jean Philippe F. Gourdine, Catherine A. Groden, Andrea L. Gropman, Melissa Haendel, Rizwan Hamid, Neil A. Hanchard, Frances High, Ingrid A. Holm, Jason Hom, Ellen M. Howerton, Yong Huang, Fariha Jamal, Yong hui Jiang, Jean M. Johnston, Angela L. Jones, Lefkothea Karaviti, David M. Koeller, Isaac S. Kohane, Jennefer N. Kohler, Donna M. Krasnewich, Susan Korrick, Mary Koziura, Joel B. Krier, Jennifer E. Kyle, Seema R. Lalani, C. Christopher Lau, Jozef Lazar, Kimberly LeBlanc, Brendan H. Lee, Hane Lee, Shawn E. Levy, Richard A. Lewis, Sharyn A. Lincoln, Sandra K. Loo, Joseph Loscalzo, Richard L. Maas, Ellen F. Macnamara, Calum A. MacRae, Valerie V. Maduro, Marta M. Majcherska, May Christine V. Malicdan, Laura A. Mamounas, Teri A. Manolio, Thomas C. Markello, Ronit Marom, Martin G. Martin, Julian A. Martínez-Agosto, Shruti Marwaha, Thomas May, Allyn McConkie-Rosell, Colleen E. McCormack, Alexa T. McCray, Jason D. Merker, Thomas O. Metz, Matthew Might, Paolo M. Moretti, Marie Morimoto, John J. Mulvihill, David R. Murdock, Jennifer L. Murphy, Donna M. Muzny, Michele E. Nehrebecky, Stan F. Nelson, J. Scott Newberry, John H. Newman, Sarah K. Nicholas, Donna Novacic, Jordan S. Orange, James P. Orengo, J. Carl Pallais, Christina Gs Palmer, Jeanette C. Papp, Neil H. Parker, Loren Dm Pena, John A. Phillips, Jennifer E. Posey, John H. Postlethwait, Lorraine Potocki, Barbara N. Pusey, Genecee Renteria, Chloe M. Reuter, Lynette Rives, Amy K. Robertson, Lance H. Rodan, Jill A. Rosenfeld, Jacinda B. Sampson, Susan L. Samson, Kelly Schoch, Daryl A. Scott, Lisa Shakachite, Prashant Sharma, Vandana Shashi, Rebecca Signer, Edwin K. Silverman, Janet S. Sinsheimer, Kevin S. Smith, Rebecca C. Spillmann, Joan M. Stoler, Nicholas Stong, Jennifer A. Sullivan, David A. Sweetser,

Queenie K.G. Tan, Cynthia J. Tifft, Camilo Toro, Alyssa A. Tran, Tiina K. Urv, Eric Vilain, Tiphanie P. Vogel, Daryl M. Waggott, Colleen E. Wahl, Nicole M. Walley, Chris A. Walsh, Melissa Walker, Jijun Wan, Michael F. Wangler, Patricia A. Ward, Katrina M. Waters, Bobbie Jo M. Webb-Robertson, Monte Westerfield, Matthew T. Wheeler, Anastasia L. Wise, Lynne A. Wolfe, Elizabeth A. Worthey, Shinya Yamamoto, John Yang, Yaping Yang, Amanda J. Yoon, Guoyun Yu, Diane B. Zastrow, Chunli Zhao, Allison Zheng, Kym Boycott, Alex MacKenzie, Jacek Majewski, Michael Brudno, Dennis Bulman, David Dyment, Lars Lind, Erik Ingelsson, Alexis Battle, Gill Bejerano, Jonathan A. Bernstein, Euan A. Ashley, Kym M. Boycott, Jason D. Merker, Matthew T. Wheeler, and Stephen B. Montgomery. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*, 25(6):911–919, 2019. ISSN 1546170X. doi: 10.1038/s41591-019-0457-8.

[28] Hernan D Gonorazky, Sergey Naumenko, Arun K Ramani, Viswateja Nelakuditi, Pouria Mashouri, Peiqui Wang, Dennis Kao, Krish Ohri, Senthuri Viththiyapaskaran, Mark A Tarnopolsky, Katherine D Mathews, Steven A Moore, Andres N Osorio, David Villanova, Dwi U Kemaladewi, Ronald D Cohn, Michael Brudno, and James J Dowling. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *The American Journal of Human Genetics*, 0(0):466–483, 2019. ISSN 00029297. doi: 10.1016/j.ajhg.2019.01.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929719300126.

[29] Revital Bronstein, Elizabeth E. Capowski, Sudeep Mehrotra, Alex D. Jansen, Daniel Navarro-Gomez, Mathew Maher, Emily Place, Riccardo Sangermano, Kinga M. Bujakowska, David M. Gamm, and Eric A. Pierce. A combined RNA-seq and whole genome sequencing approach for identification of non-coding pathogenic variants in single families. *bioRxiv*, page 766717, 2019. doi: 10.1101/766717. URL https://www.biorxiv.org/content/10.1101/766717v1.

[30] Joseph K Aicher, Paul Jewell, Jorge Vaquero-Garcia, Yoseph Barash, and Elizabeth J Bhoj. Mapping RNA splicing variations in clinically-accessible and non-accessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *bioRxiv*, page 727586, 2019. doi: 10.1101/727586. URL https://www.biorxiv.org/content/10.1101/727586v1.

[31] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014. ISSN 1474760X. doi: 10.1186/s13059-014-0550-8.

[32] Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson, Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, 2018. ISSN 15461718. doi: 10.1038/s41588-017-0004-9. URL http://dx.doi.org/10.1038/s41588-017-0004-9.

[33] Felix Brechtmann, Christian Mertes, Agnė Matusevičiūtė, Vicente A Yépez, Žiga Avsec, Maximilian Herzog, Daniel M Bader, Holger Prokisch, and Julien Gagneur. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *American Journal of Human Genetics*, 103(6):907–917, 2018. ISSN 15376605. doi: 10.1016/j.ajhg.2018.10.025.

[34] Pejman Mohammadi, Stephane E. Castel, Beryl B. Cummings, Jonah Einson, Christina Sousa, Paul Hoffman, Sandra Donkervoort, Payam Mohassel, Reghan Foley, Heather E. Wheeler, Hae Kyung Im, Carsten G. Bonnemann, Daniel G. MacArthur, and Tuuli Lappalainen. Quantifying genetic regulatory variation in human populations improves transcriptome analysis in rare disease patients. *bioRxiv*, page 632794, 2019. doi: 10.1101/632794. URL https://www.biorxiv.org/content/10.1101/632794v1.

[35] Garrett Jenkinson, Yang I Li, Shubham Basu, Margot A Cousin, Gavin R Oliver, and Eric W Klee. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics*, pages 1–7, apr 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/

btaa259. URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa259/5823301.

[36] Christian Mertes, Ines F. Scheller, Vicente A. Yépez, Muhammed H. Çelik, Yingjiqiong Liang, Laura S. Kremer, Mirjana Gusic, Holger Prokisch, and Julien Gagneur. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nature Communications*, 12(1):1–13, 2021. ISSN 20411723. doi: 10.1038/s41467-020-20573-7. URL http://dx.doi.org/10.1038/s41467-020-20573-7.

[37] Matthew K Iyer, Yashar S Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R Barrette, John R Prensner, Joseph R Evans, Shuang Zhao, Anton Poliakov, Xuhong Cao, Saravana M Dhanasekaran, Yi Mi Wu, Dan R Robinson, David G Beer, Felix Y Feng, Hariharan K Iyer, and Arul M Chinnaiyan. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, 47(3):199–208, 2015. ISSN 15461718. doi: 10.1038/ng.3192.

[38] Marina M. Scotti and Maurice S. Swanson. RNA mis-splicing in disease, 2016. ISSN 14710064. URL http://dx.doi.org/10.1038/nrg.2015.3.

[39] Benjamin J. Blencowe. Alternative Splicing: New Insights from Global Analyses. *Cell*, 126(1):37–47, 2006. ISSN 00928674. doi: 10.1016/j.cell.2006.06.023.

[40] Yang I Li, Bryce Van De Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016. ISSN 10959203. doi: 10.1126/science.aad9417.

[41] Rachel Soemedi, Kamil J. Cygan, Christy L. Rhine, Jing Wang, Charlston Bulacan, John Yang, Pinar Bayrak-Toydemir, Jamie McDonald, and William G. Fairbrother. Pathogenic variants that alter protein code often disrupt splicing. *Nature Genetics*, 49(6):848–855, 2017. ISSN 15461718. doi: 10.1038/ng.3837. URL http://dx.doi.org/10.1038/ng.3837.

[42] Leonardo Collado-Torres, Abhinav Nellore, and Andrew E. Jaffe. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Research*, 6(0):1558, aug 2017. ISSN 2046-1402. doi: 10.12688/f1000research.12223.1. URL https://f1000research.com/articles/6-1558/v1.

[43] Htoo Wai, Andrew G.L. Douglas, and Diana Baralle. RNA splicing analysis in genomic medicine. *International Journal of Biochemistry and Cell Biology*, 108(December 2018): 61–71, 2019. ISSN 18785875. doi: 10.1016/j.biocel.2018.12.009. URL https://doi.org/10.1016/j.biocel.2018.12.009.

[44] Jinkuk Kim, Chunguang Hu, Christelle Moufawad El Achkar, Lauren E. Black, Julie Douville, Austin Larson, Mary K. Pendergast, Sara F. Goldkind, Eunjung A. Lee, Ashley Kuniholm, Aubrie Soucy, Jai Vaze, Nandkishore R. Belur, Kristina Fredriksen, Iva Stojkovska, Alla Tsytsykova, Myriam Armant, Renata L. DiDonato, Jaejoon Choi, Laura Cornelissen, Luis M. Pereira, Erika F. Augustine, Casie A. Genetti, Kira Dies, Brenda Barton, Lucinda Williams, Benjamin D. Goodlett, Bobbie L. Riley, Amy Pasternak, Emily R. Berry, Kelly A. Pflock, Stephen Chu, Chantal Reed, Kimberly Tyndall, Pankaj B. Agrawal, Alan H. Beggs, P. Ellen Grant, David K. Urion, Richard O. Snyder, Susan E. Waisbren, Annapurna Poduri, Peter J. Park, Al Patterson, Alessandra Biffi, Joseph R. Mazzulli, Olaf Bodamer, Charles B. Berde, and Timothy W. Yu. Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. *New England Journal of Medicine*, 381(17):1644–1652, 2019. ISSN 0028-4793. doi: 10.1056/nejmoa1813279.

[45] Christopher R. Sibley, Warren Emmett, Lorea Blazquez, Ana Faro, Nejc Haberman, Michael Briese, Daniah Trabzuni, Mina Ryten, Michael E. Weale, John Hardy, Miha Modic, Tomaž Curk, Stephen W. Wilson, Vincent Plagnol, and Jernej Ule. Recursive splicing in long vertebrate genes. *Nature*, 2015. ISSN 14764687. doi: 10.1038/nature14466.

[46] Gene Yeo, Dirk Holste, Gabriel Kreiman, and Christopher B. Burge. Variation in alternative splicing across human tissues. *Genome biology*, 5(10):1–15, 2004. ISSN 14656914. doi: 10.1186/gb-2004-5-10-r74.

[47] Yong E. Zhang, Patrick Landback, Maria Vibranovski, and Manyuan Long. New genes expressed in human brains: Implications for annotating evolving genomes. *BioEssays*, 34 (11):982–991, 2012. ISSN 02659247. doi: 10.1002/bies.201200008.

[48] Tamara Steijger, Josep F. Abril, Pär G. Engström, Felix Kokocinski, Martin Akerman, Tyler Alioto, Giovanna Ambrosini, Stylianos E. Antonarakis, Jonas Behr, Paul Bertone, Regina Bohnert, Philipp Bucher, Nicole Cloonan, Thomas Derrien, Sarah Djebali, Jiang Du, Sandrine Dudoit, Mark Gerstein, Thomas R. Gingeras, David Gonzalez, Sean M. Grimmond, Roderic Guigó, Lukas Habegger, Jennifer Harrow, Tim J. Hubbard, Christian Iseli, Géraldine Jean, André Kahles, Julien Lagarde, Jing Leng, Gregory Lefebvre, Suzanna Lewis, Ali Mortazavi, Peter Niermann, Gunnar Rätsch, Alexandre Reymond, Paolo Ribeca, Hugues Richard, Jacques Rougemont, Joel Rozowsky, Michael Sammeth, Andrea Sboner, Marcel H. Schulz, Steven M.J. Searle, Naryttza Diaz Solorzano, Victor Solovyev, Mario Stanke, Brian J. Stevenson, Heinz Stockinger, Armand Valsesia, David Weese, Simon White, Barbara J. Wold, Jie Wu, Thomas D. Wu, Georg Zeller, Daniel Zerbino, and Michael Q. Zhang. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, 10(12):1177–1184, 2013. ISSN 15487091. doi: 10.1038/nmeth.2714.

[49] Irwin Jungreis, Michael L. Tress, Jonathan Mudge, Cristina Sisu, and Toby Hunt. Nearly all new protein-coding predictions in the CHESS database are not protein-coding. *bioRxiv*, pages 1–19, 2018. doi: 10.1101/360602. URL https://www.biorxiv.org/content/early/2018/07/02/360602.

[50] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(DATABASE ISS.):514–517, 2005. ISSN 03051048. doi: 10.1093/nar/gki033.

[51] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, 348(6235):648–60, may 2015. ISSN 1095-9203. doi: 10.1126/science.1262110. URL http://www.ncbi.nlm.nih.gov/pubmed/25954001%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4547484http://www.sciencemag.org/cgi/doi/10.1126/science.1262110http://www.ncbi.nlm.nih.gov/pubmed/25954001http://www.pubmedcentral.nih.gov/artic.

[52] Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Reproducible RNA-seq analysis using recount2. *Nature Biotechnology*, 35(4):319–321, apr 2017. ISSN 1087-0156. doi: 10.1038/nbt.3838. URL http://www.nature.com/doifinder/10.1038/nbt.3838http://www.nature.com/articles/nbt.3838.

[53] Leonardo Collado-Torres, Abhinav Nellore, Alyssa C. Frazee, Christopher Wilks, Michael I. Love, Ben Langmead, Rafael A. Irizarry, Jeffrey T. Leek, and Andrew E. Jaffe. Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Research*, 45(2):e9, 2017. ISSN 13624962. doi: 10.1093/nar/gkw852.

[54] Manuel Irimia, Robert J Weatheritt, Jonathan D Ellis, Neelroop N Parikshak, Thomas Gonatopoulos-Pournatzis, Mariana Babor, Mathieu Quesnel-Vallières, Javier Tapial, Bushra Raj, Dave O'Hanlon, Miriam Barrios-Rodiles, Michael J E Sternberg, Sabine P Cordes, Frederick P Roth, Jeffrey L Wrana, Daniel H Geschwind, and Benjamin J Blencowe. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, 159(7):1511–1523, 2014. ISSN 10974172. doi: 10.1016/j.cell.2014.11.035.

[55] Adam Labadorf, Andrew G. Hoss, Valentina Lagomarsino, Jeanne C. Latourelle, Tiffany C. Hadzi, Joli Bregu, Marcy E. MacDonald, James F. Gusella, Jiang Fan Chen, Schahram Akbarian, Zhiping Weng, and Richard H. Myers. RNA sequence analysis of human

huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression. *PLoS ONE*, 10(12):1–21, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0143563.

[56] J. W. Thomas, J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, A. C. Siepel, P. J. Thomas, J. C. McDowell, B. Maskeri, N. F. Hansen, M. S. Schwartz, R. J. Weber, W. J. Kent, D. Karolchik, T. C. Bruen, R. Bevan, D. J. Cutler, S. Schwartz, L. Elnitski, J. R. Idol, A. B. Prasad, S. Q. Lee-Lin, V. V.B. Maduro, T. J. Summers, M. E. Portnoy, N. L. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Cariaga, C. P. Brinkley, S. Y. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghighi, S. L. Ho, M. C. Huang, E. Karlins, P. L. Laric, R. Legaspi, M. J. Lim, Q. L. Maduro, C. A. Masiello, S. D. Mastrian, J. C. McCloskey, R. Pearson, S. Stantripop, E. E. Tiongson, J. T. Tran, C. Tsurgeon, J. L. Vogt, M. A. Walker, K. D. Wetherby, L. S. Wiggins, A. C. Young, L. H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C. L. Shu, P. J. De Jong, C. E. Lawrence, A. F. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E. D. Green. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 2003. ISSN 00280836. doi: 10.1038/nature01858.

[57] Julia Di Iulio, Istvan Bartha, Emily H.M. Wong, Hung Chun Yu, Victor Lavrenko, Dongchan Yang, Inkyung Jung, Michael A. Hicks, Naisha Shah, Ewen F. Kirkness, Martin M. Fabani, William H. Biggs, Bing Ren, J. Craig Venter, and Amalio Telenti. The human noncoding genome defined by genetic diversity. *Nature Genetics*, 50(3):333–337, 2018. ISSN 15461718. doi: 10.1038/s41588-018-0062-7. URL http://dx.doi.org/10.1038/s41588-018-0062-7.

[58] Nathan G. Skene and Seth G.N. Grant. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Frontiers in Neuroscience*, 10(JAN):1–11, 2016. ISSN 1662453X. doi: 10.3389/fnins.2016.00016.

[59] Christopher Wilks, Omar Ahmed, Daniel N Baker, David Zhang, Leonardo Collado-Torres, and Ben Langmead. Megadepth: efficient coverage quantification for BigWigs and BAMs. *Bioinformatics*, (March):1–3, 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab152.

[60] Shanrong Zhao, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific Reports*, 8(1):4781, dec 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-23226-4. URL http://dx.doi.org/10.1038/s41598-018-23226-4http://www.nature.com/articles/s41598-018-23226-4.

[61] W Ford Doolittle. We simply cannot go on being so vague about ’function’. *Genome Biology*, 19(1):18–20, 2018. ISSN 1474760X. doi: 10.1186/s13059-018-1600-4.

[62] Gaia Novarino, Ali G. Fenstermaker, Maha S. Zaki, Matan Hofree, Jennifer L. Silhavy, Andrew D. Heiberg, Mostafa Abdellateef, Basak Rosti, Eric Scott, Lobna Mansour, Amira Masri, Hulya Kayserili, Jumana Y. Al-Aama, Ghada M.H. Abdel-Salam, Ariana Karminejad, Majdi Kara, Bulent Kara, Bita Bozorgmehri, Tawfeg Ben-Omran, Faezeh Mojahedi, Iman Gamal El Din Mahmoud, Naima Bouslam, Ahmed Bouhouche, Ali Benomar, Sylvain Hanein, Laure Raymond, Sylvie Forlani, Massimo Mascaro, Laila Selim, Nabil Shehata, Nasir Al-Allawi, P. S. Bindu, Matloob Azam, Murat Gunel, Ahmet Caglayan, Kaya Bilguvar, Aslihan Tolun, Mahmoud Y. Issa, Jana Schroth, Emily G. Spencer, Rasim O. Rosti, Naiara Akizu, Keith K. Vaux, Anide Johansen, Alice A. Koh, Hisham Megahed, Alexandra Durr, Alexis Brice, Giovanni Stevanin, Stacy B. Gabriel, Trey Ideker, and Joseph G. Gleeson. Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. *Science*, 343(6170):506–511, 2014. ISSN 10959203. doi: 10.1126/science.1247363.

[63] Christopher Wilks, Leonardo Collado-Torres, S C. Zheng, Andrew E. Jaffe, Abhinav Nellore, Kasper D. Hansen, and Ben Langmead. Explore and download data from the

recount3 project. *R package version 1.0.7*, 2020. doi: https://doi.org/doi:10.18129/B9.bioc.recount3. URL https://doi.org/doi:10.18129/B9.bioc.recount3.

[64] Siddharth Sethi, David Zhang, Sebastian Guelfi, and Zhongbo Chen. Leveraging omic features with F3UTER enables identification of unannotated 3 ' UTRs for synaptic genes. pages 1–44, 2021.

[65] Miten Jain, Hugh E. Olsen, Benedict Paten, and Mark Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):1–11, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1103-0. URL http://dx.doi.org/10.1186/s13059-016-1103-0.

[66] Glennis A. Logsdon, Mitchell R. Vollger, and Evan E. Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, 2020. ISSN 14710064. doi: 10.1038/s41576-020-0236-x. URL http://dx.doi.org/10.1038/s41576-020-0236-x.

[67] Prech Uapinyoying, Jeremy Goecks, Susan M. Knoblac, Karuna Panchapakesan, Carsten G. Bonneman, Terence A. Partridg, Jyoti K. Jaiswa, and Eric P. Hoffma. A long-read RNA-seq approach to identify novel transcripts of very large genes. *Genome Research*, 30(6):885–897, 2020. ISSN 15495469. doi: 10.1101/gr.259903.119.

[68] Christian R. Schubert, Patricio O'Donnell, Jie Quan, Jens R. Wendland, Hualin S. Xi, Ashley R. Winslow, Enrico Domenici, Laurent Essioux, Tony Kam-Thong, David C. Airey, John N. Calley, David A. Collier, Hong Wang, Brian Eastwood, Philip Ebert, Yushi Liu, Laura Nisenbaum, Cara Ruble, James Scherschel, Ryan Matthew Smith, Hui Rong Qian, Kalpana Merchant, Michael Didriksen, Mitsuyuki Matsumoto, Takeshi Saito, Nicholas J. Brandon, Alan J. Cross, Qi Wang, Husseini Manji, Hartmuth Kolb, Maura Furey, Wayne C. Drevets, Joo Heon Shin, Andrew E. Jaffe, Yankai Jia, Richard E. Straub, Amy Deep-Soboslay, Thomas M. Hyde, Joel E. Kleinman, and Daniel R. Weinberger. BrainSeq: Neurogenomics to Drive Novel Target Discovery for Neuropsychiatric Disorders. *Neuron*, 88(6):1078–1083, 2015. ISSN 10974199. doi: 10.1016/j.neuron.2015.10.047.

[69] Daifeng Wang, Shuang Liu, Jonathan Warrell, Hyejung Won, Xu Shi, Fabio C.P. Navarro, Declan Clarke, Mengting Gu, Prashant Emani, Yucheng T. Yang, X. Min, Michael J. Gandal, Shaoke Lou, Jing Zhang, Jonathan J. Park, Chengfei Yan, Suhn KyongRhie, Kasidet Manakongtreecheep, Holly Zhou, A. Aparna Natha, Mette Peters, Eugenio Mattei, Dominic Fitzgerald, Tonya Brunetti, Jill Moore, Yan Jiang, Kiran Girdhar, Gabriel E. Hoffman, Selim Kalayci, Zeynep H. Gümüş, Gregory E. Crawford, Panos Roussos, Schahram Akbarian, Andrew E. Jaffe, Kevin P. White, Zhiping Weng, Nenad Sestan, Daniel H. Geschwind, James A. Knowles, and Mark B. Gerstein. Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362(6420), 2018. ISSN 10959203. doi: 10.1126/science.aat8464.

[70] Muzlifah Haniffa, Deanne Taylor, Sten Linnarsson, Bruce J. Aronow, Gary D. Bader, Roger A. Barker, Pablo G. Camara, J. Gray Camp, Alain Chédotal, Andrew Copp, Heather C. Etchevers, Paolo Giacobini, Berthold Göttgens, Guoji Guo, Ania Hupalowska, Kylie R. James, Emily Kirby, Arnold Kriegstein, Joakim Lundeberg, John C. Marioni, Kerstin B. Meyer, Kathy K. Niakan, Mats Nilsson, Bayanne Olabi, Dana Pe'er, Aviv Regev, Jennifer Rood, Orit Rozenblatt-Rosen, Rahul Satija, Sarah A. Teichmann, Barbara Treutlein, Roser Vento-Tormo, Simone Webb, Pascal Barbry, Omer Bayraktar, Sam Behjati, Andreas Bosio, Bruno Canque, Frédéric Chalmel, Yorick Gitton, Deborah Henderson, Anne Jorgensen, Steven Lisgo, Jinyue Liu, Emma Lundberg, Jean Léon Maitre, Séverine Mazaud-Guittot, Elizabeth Robertson, Antoine Rolland, Raphael Scharfmann, Michèle Souyri, Erik Sundström, Stéphane Zaffran, and Matthias Zilbauer. A roadmap for the Human Developmental Cell Atlas. *Nature*, 597(7875):196–205, 2021. ISSN 14764687. doi: 10.1038/s41586-021-03620-1.

[71] Kym M Boycott, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation, 2013. ISSN 14710056. URL http://dx.doi.org/10.1038/nrg3555.

[72] Janine Meienberg, Rémy Bruggmann, Konrad Oexle, and Gabor Matyas. Clinical sequencing: is WGS the better WES? *Human Genetics*, 135(3):359–362, 2016. ISSN 14321203. doi: 10.1007/s00439-015-1631-9.

[73] Charles A. Steward, Alasdair P.J. Parker, Berge A. Minassian, Sanjay M. Sisodiya, Adam Frankish, and Jennifer Harrow. Genome annotation for clinical genomic diagnostics: Strengths and weaknesses, 2017. ISSN 1756994X.

[74] Jian Wang and Yiping Shen. When a "disease-causing mutation" is not a pathogenic variant. *Clinical Chemistry*, 60(5):711–713, 2014. ISSN 15308561. doi: 10.1373/clinchem.2013.215947.

[75] Karen Eilbeck, Aaron Quinlan, and Mark Yandell. Settling the score: Variant prioritization and Mendelian disease. *Nature Reviews Genetics*, 18(10):599–612, 2017. ISSN 14710064. doi: 10.1038/nrg.2017.52. URL http://dx.doi.org/10.1038/nrg.2017.52.

[76] Jenny Lord, Giuseppe Gallone, Patrick J. Short, Jeremy F. McRae, Holly Ironfield, Elizabeth H. Wynn, Sebastian S. Gerety, Liu He, Bronwyn Kerr, Diana S. Johnson, Emma McCann, Esther Kinning, Frances Flinter, I. Karen Temple, Jill Clayton-Smith, Meriel McEntagart, Sally Ann Lynch, Shelagh Joss, Sofia Douzgou, Tabib Dabir, Virginia Clowes, Vivienne P.M. McConnell, Wayne Lam, Caroline F. Wright, David R. FitzPatrick, Helen V. Firth, Jeffrey C. Barrett, and Matthew E. Hurles. Pathogenicity and selective constraint on variation near splice sites. *Genome Research*, pages 159–170, 2019. doi: 10.1101/256636.

[77] Eddie Park, Zhicheng Pan, Zijun Zhang, Lan Lin, and Yi Xing. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *American Journal of Human Genetics*, 102(1):11–26, 2018. ISSN 15376605. doi: 10.1016/j.ajhg.2017.11.002. URL https://doi.org/10.1016/j.ajhg.2017.11.002.

[78] Xin Li, Yungil Kim, Emily K. Tsang, Joe R. Davis, Farhan N. Damani, Colby Chiang, Gaelen T. Hess, Zachary Zappala, Benjamin J. Strober, Alexandra J. Scott, Amy Li, Andrea Ganna, Michael C. Bassik, Jason D. Merker, François Aguet, Kristin G. Ardlie, Beryl B. Cummings, Ellen T. Gelfand, Gad Getz, Kane Hadley, Robert E. Handsaker, Katherine H. Huang, Seva Kashin, Konrad J. Karczewski, Monkol Lek, Xiao Li, Daniel G. MacArthur, Jared L. Nedzel, Duyen T. Nguyen, Michael S. Noble, Ayellet V. Segrè, Casandra A. Trowbridge, Taru Tukiainen, Nathan S. Abell, Brunilda Balliu, Ruth Barshir, Omer Basha, Gireesh K. Bogu, Andrew Brown, Christopher D. Brown, Stephane E. Castel, Lin S. Chen, Donald F. Conrad, Nancy J. Cox, Olivier Delaneau, Emmanouil T. Dermitzakis, Barbara E. Engelhardt, Eleazar Eskin, Pedro G. Ferreira, Laure Frésard, Eric R. Gamazon, Diego Garrido-Martín, Ariel D.H. Gewirtz, Genna Gliner, Michael J. Gloudemans, Roderic Guigo, Ira M. Hall, Buhm Han, Yuan He, Farhad Hormozdiari, Cedric Howald, Hae Kyung Im, Brian Jo, Eun Yong Kang, Sarah Kim-Hellmuth, Tuuli Lappalainen, Gen Li, Boxiang Liu, Serghei Mangul, Mark I. McCarthy, Ian C. McDowell, Pejman Mohammadi, Jean Monlong, Manuel Muñoz-Aguirre, Anne W. Ndungu, Dan L. Nicolae, Andrew B. Nobel, Meritxell Oliva, Halit Ongen, John J. Palowitch, Nikolaos Panousis, Panagiotis Papasaikas, Yoson Park, Princy Parsana, Anthony J. Payne, Christine B. Peterson, Jie Quan, Ferran Reverter, Chiara Sabatti, Ashis Saha, Michael Sammeth, Andrey A. Shabalin, Reza Sodaei, Matthew Stephens, Barbara E. Stranger, Jae Hoon Sul, Sarah Urbut, Martijn Van De Bunt, Gao Wang, Xiaoquan Wen, Fred A. Wright, Hualin S. Xi, Esti Yeger-Lotem, Judith B. Zaugg, Yi Hui Zhou, Joshua M. Akey, Daniel Bates, Joanne Chan, Melina Claussnitzer, Kathryn Demanelis, Morgan Diegel, Jennifer A. Doherty, Andrew P. Feinberg, Marian S. Fernando, Jessica Halow, Kasper D. Hansen, Eric Haugen, Peter F. Hickey, Lei Hou, Farzana Jasmine, Ruiqi Jian, Lihua Jiang, Audra Johnson, Rajinder Kaul, Manolis Kellis, Muhammad G. Kibriya, Kristen Lee, Jin Billy Li, Qin Li, Jessica Lin, Shin Lin, Sandra Linder, Caroline Linke, Yaping Liu, Matthew T. Maurano, Benoit Molinie,

Jemma Nelson, Fidencio J. Neri, Yongjin Park, Brandon L. Pierce, Nicola J. Rinaldi, Lindsay F. Rizzardi, Richard Sandstrom, Andrew Skol, Kevin S. Smith, Michael P. Snyder, John Stamatoyannopoulos, Hua Tang, Li Wang, Meng Wang, Nicholas Van Wittenberghe, Fan Wu, Rui Zhang, Concepcion R. Nierras, Philip A. Branton, Latarsha J. Carithers, Ping Guan, Helen M. Moore, Abhi Rao, Jimmie B. Vaught, Sarah E. Gould, Nicole C. Lockart, Casey Martin, Jeffery P. Struewing, Simona Volpi, Anjene M. Addington, Susan E. Koester, A. Roger Little, Lori E. Brigham, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Gene Kopen, William F. Leinweber, John T. Lonsdale, Alisa McDonald, Bernadette Mestichelli, Kevin Myer, Brian Roe, Michael Salvatore, Saboor Shad, Jeffrey A. Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Jason Bridge, Barbara A. Foster, Bryan M. Gillard, Ellen Karasik, Rachna Kumar, Mark Miklos, Michael T. Moser, Scott D. Jewell, Robert G. Montroy, Daniel C. Rohrer, Dana R. Valley, David A. Davis, Deborah C. Mash, Anita H. Undale, Anna M. Smith, David E. Tabor, Nancy V. Roche, Jeffrey A. McLean, Negin Vatanian, Karna L. Robinson, Leslie Sobin, Mary E. Barcus, Kimberly M. Valentino, Liqun Qi, Steven Hunter, Pushpa Hariharan, Shilpi Singh, Ki Sung Um, Takunda Matose, Maria M. Tomaszewski, Laura K. Barker, Maghboeba Mosavel, Laura A. Siminoff, Heather M. Traino, Paul Flicek, Thomas Juettemann, Magali Ruffier, Dan Sheppard, Kieron Taylor, Stephen J. Trevanion, Daniel R. Zerbino, Brian Craft, Mary Goldman, Maximilian Haeussler, W. James Kent, Christopher M. Lee, Benedict Paten, Kate R. Rosenbloom, John Vivian, Jingchun Zhu, Alexis Battle, and Stephen B. Montgomery. The impact of rare variation on gene expression across tissues. *Nature*, 550(7675):239–243, 2017. ISSN 14764687. doi: 10.1038/nature24267. URL http://dx.doi.org/10.1038/nature24267.

[79] Jorge Vaquero-Garcia, Alejandro Barrera, Matthew R. Gazzara, Juan Gonzalez-Vallinas, Nicholas F. Lahens, John B. Hogenesch, Kristen W. Lynch, and Yoseph Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5(FEBRUARY2016):1–30, 2016. ISSN 2050084X. doi: 10.7554/eLife.11752.

[80] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884—-i890, 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty560.

[81] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. ISSN 13674803. doi: 10.1093/bioinformatics/bts635.

[82] Liguo Wang, Shengqin Wang, and Wei Li. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics*, 28(16):2184–2185, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts356.

[83] Leonardo Collado-Torres, Abhinav Nellore, Alyssa C. Frazee, Christopher Wilks, Michael I. Love, Ben Langmead, Rafael A. Irizarry, Jeffrey T. Leek, and Andrew E. Jaffe. Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Research*, 45(2):e9, 2017. ISSN 13624962. doi: 10.1093/nar/gkw852.

[84] Christopher Wilks, Phani Gaddipati, Abhinav Nellore, and Ben Langmead. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics (Oxford, England)*, 34(1):114–116, 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx547. URL http://www.ncbi.nlm.nih.gov/pubmed/28968689http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5870547.

[85] Vicente A. Yépez, Christian Mertes, Michaela F. Müller, Daniela Klaproth-Andrade, Leonhard Wachutka, Laure Frésard, Mirjana Gusic, Ines F. Scheller, Patricia F. Goldberg, Holger Prokisch, and Julien Gagneur. Detection of aberrant gene expression events in RNA sequencing data. *Nature Protocols*, 16(2):1276–1296, 2021. ISSN 17502799. doi: 10.1038/s41596-020-00462-5. URL http://dx.doi.org/10.1038/s41596-020-00462-5.

[86] Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, 9(1):1–5, 2019. ISSN 20452322. doi: 10.1038/s41598-019-45839-z.

[87] Robert D S Pitceathly, Shamima Rahman, Yehani Wedatilake, James M Polke, Sebahattin Cirak, A Reghan Foley, Anna Sailer, Matthew E Hurles, Jim Stalker, Iain Hargreaves, Cathy E Woodward, Mary G Sweeney, Francesco Muntoni, Henry Houlden, Jan Willem Taanman, and Michael G Hanna. NDUFA4 Mutations Underlie Dysfunction of a Cytochrome c Oxidase Subunit Linked to Human Neurological Disease. *Cell Reports*, 3(6):1795–1805, 2013. ISSN 22111247. doi: 10.1016/j.celrep.2013.05.005. URL http://dx.doi.org/10.1016/j.celrep.2013.05.005.

[88] M. G.P. van der Wijst, D. H. de Vries, H. E. Groot, G. Trynka, C. C. Hon, M. J. Bonder, O. Stegle, M. C. Nawijn, Y. Idaghdour, P. van der Harst, C. J. Ye, J. Powell, F. J. Theis, A. Mahfouz, M. Heinig, and L. Franke. The single-cell eQTLGen consortium. *eLife*, 9: 1–21, 2020. ISSN 2050084X. doi: 10.7554/eLife.52155.

[89] Rachel Soemedi, Kamil J. Cygan, Christy L. Rhine, Jing Wang, Charlston Bulacan, John Yang, Pinar Bayrak-Toydemir, Jamie McDonald, and William G. Fairbrother. Pathogenic variants that alter protein code often disrupt splicing. *Nature Genetics*, 49(6):848–855, 2017. ISSN 15461718. doi: 10.1038/ng.3837. URL http://dx.doi.org/10.1038/ng.3837.

[90] Xueqiu Jian, Eric Boerwinkle, and Xiaoming Liu. In silico tools for splicing defect prediction: A survey from the viewpoint of end users. *Genetics in Medicine*, 16(7): 497–503, 2014. ISSN 15300366. doi: 10.1038/gim.2013.176.

[91] Jennifer E. Posey, Tamar Harel, Pengfei Liu, Jill A. Rosenfeld, Regis A. James, Zeynep H. Coban Akdemir, Magdalena Walkiewicz, Weimin Bi, Rui Xiao, Yan Ding, Fan Xia, Arthur L. Beaudet, Donna M. Muzny, Richard A. Gibbs, Eric Boerwinkle, Christine M. Eng, V. Reid Sutton, Chad A. Shaw, Sharon E. Plon, Yaping Yang, and James R. Lupski. Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *New England Journal of Medicine*, 376(1):21–31, 2017. ISSN 0028-4793. doi: 10.1056/nejmoa1516767.

[92] Ambroise Wonkam, Emile R. Chimusa, Khuthala Mnika, Gift Dineo Pule, Valentina Josiane Ngo Bitoungui, Nicola Mulder, Daniel Shriner, Charles N. Rotimi, and Adebowale Adeyemo. Genetic modifiers of long-term survival in sickle cell anemia. *Clinical and Translational Medicine*, 10(4):1–15, 2020. ISSN 2001-1326. doi: 10.1002/ctm2.152.

[93] Michael S.D. Kormann, Alexander Dewerth, Felizitas Eichner, Praveen Baskaran, Andreas Hector, Nicolas Regamey, Dominik Hartl, Rupert Handgretinger, and Justin S. Antony. Transcriptomic profile of cystic fibrosis patients identifies type I interferon response and ribosomal stalk proteins as potential modifiers of disease severity. *PLoS ONE*, 12(8):1–13, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0183526.

[94] Juan A Botia, Sebastian Guelfi, David Zhang, Karishma D039;Sa, Regina Reinolds, Daniel Onah, Ellen M Mcdonagh, Antonio Rueda-Martin, Arianna Tucci, Augusto Rendon, Henry Houlden, John Hardy, Mina Ryten, Juan A Botía, Sebastian Guelfi, David Zhang, Karishma D Sa, Regina Reynolds, Ellen M Mcdonagh, Antonio Rueda Martin, Arianna Tucci, Augusto Rendon, Juan A Botia, Sebastian Guelfi, David Zhang, Karishma D039;Sa, Regina Reinolds, Daniel Onah, Ellen M Mcdonagh, Antonio Rueda-Martin, Arianna Tucci, Augusto Rendon, Henry Houlden, John Hardy, and Mina Ryten. G2P: Using machine learning to understand and predict genes causing rare neurological disorders. *bioRxiv*, pages 1–38, 2018. doi: http://dx.doi.org/10.1101/288845. URL http://biorxiv.org/content/early/2018/03/27/288845.abstract.

[95] Jack J. Collier, Claire Guissart, Monika Oláhová, Souphatta Sasorith, Florence Piron-Prunier, Fumi Suomi, David Zhang, Nuria Martinez-Lopez, Nicolas Leboucq, Angela Bahr, Silvia Azzarello-Burri, Selina Reich, Ludger Schöls, Tuomo M. Polvikoski, Pierre

Meyer, Lise Larrieu, Andrew M. Schaefer, Hessa S. Alsaif, Suad Alyamani, Stephan Zuchner, Inês A. Barbosa, Charu Deshpande, Angela Pyle, Anita Rauch, Matthis Synofzik, Fowzan S. Alkuraya, François Rivier, Mina Ryten, Robert McFarland, Agnès Delahodde, Thomas G. McWilliams, Michel Koenig, and Robert W. Taylor. Developmental Consequences of Defective ATG7-Mediated Autophagy in Humans. *New England Journal of Medicine*, 384(25):2406–2417, 2021. ISSN 0028-4793. doi: 10.1056/nejmoa1915722.

[96] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. *Icdm*, 2008. URL https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest.

[97] David S. Deluca, Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell, Marc Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11): 1530–1532, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts196.

[98] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(DATABASE ISS.):514–517, 2005. ISSN 03051048. doi: 10.1093/nar/gki033.

[99] Sneha Rath, Rohit Sharma, Rahul Gupta, Tslil Ast, Connie Chan, Timothy J. Durham, Russell P. Goodman, Zenon Grabarek, Mary E. Haas, Wendy H.W. Hung, Pallavi R. Joshi, Alexis A. Jourdain, Sharon H. Kim, Anna V. Kotrys, Stephanie S. Lam, Jason G. McCoy, Joshua D. Meisel, Maria Miranda, Apekshya Panda, Anupam Patgiri, Robert Rogers, Shayan Sadre, Hardik Shah, Owen S. Skinner, Tsz Leung To, Melissa A. Walker, Hong Wang, Patrick S. Ward, Jordan Wengrod, Chen Ching Yuan, Sarah E. Calvo, and Vamsi K. Mootha. MitoCarta3.0: An updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Research*, 49(D1):D1541–D1547, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa1011.

[100] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 2020. ISSN 13624962. doi: 10.1093/nar/gkz1021.

[101] H Pagès, P Aboyoun, R Gentleman, and S DebRoy. Biostrings: Efficient manipulation of biological strings, 2017.

[102] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–50, aug 2005. ISSN 1088-9051. doi: 10.1101/gr.3715005. URL http://www.genome.org/cgi/doi/10.1101/gr.3715005http://www.ncbi.nlm.nih.gov/pubmed/16024819http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1182216.

[103] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp352.

[104] Pauline C. Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003. ISSN 03051048. doi: 10.1093/nar/gkg509.

[105] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16): 1–7, 2010. ISSN 03051048. doi: 10.1093/nar/gkq603.

[106] Wenqing Fu, Timothy D. O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M. Leal, Stacey Gabriel, David Altshuler, Jay Shendure, Deborah A. Nickerson, Michael J. Bamshad, and Joshua M. Akey. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, 2013. ISSN 00280836. doi: 10.1038/nature11690.

[107] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie Laure Yaspo, Lucinda Fulton, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Aniko Sabo, Zhuoyi Huang, Lachlan J.M. Coin, Lin Fang, Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Erik P. Garrison, Deniz Kural, Wan Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, Eric Banks, Gaurav Bhatia, Guillermo Del Angel, Giulio Genovese, Heng Li, Seva Kashin, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Kathryn Beal, Avik Datta, Javier Herrero, Graham R.S. Ritchie, Daniel Zerbino, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Herwig, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver, Jeffrey A. Rosenfeld, Carlos D. Bustamante, Stephen B. Montgomery, Francisco M. De La Vega, Jake K. Byrnes, Andrew W. Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin, Yael Baran, Eliza Cerveira, Jaeho Hwang, Ankit Malhotra, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, Fiona C.L. Hyland, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Madhusudan Gujral, Amina Noor, Kenny Ye, Esteban G. Burchard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman,

W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Scott E. Devine, Hyun Min Kang, Jeffrey M. Kidd, Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla, Alan Hodgkinson, Yun Li, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Jonathan L. Marchini, Simon Myers, Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk, Yunxin Fu, Xiaoming Liu, Momiao Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Brian L. Browning, Sharon R. Browning, Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana, Chris Tyler-Smith, Cornelis A. Albers, Qasim Ayub, Yuan Chen, Vincenza Colonna, Luke Jostins, Klaudia Walter, Yali Xue, Mark B. Gerstein, Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yao Fu, Arif O. Harmanci, Mike Jin, Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Zhang, Chris Hartl, Khalid Shakir, Jeremiah Degenhardt, Sascha Meiers, Benjamin Raeder, Francesco Paolo Casale, Oliver Stegle, Eric Wubbo Lameijer, Ira Hall, Vineet Bafna, Jacob Michaelson, Eugene J. Gardner, Ryan E. Mills, Gargi Dayama, Ken Chen, Xian Fan, Zechen Chong, Tenghui Chen, Mark J. Chaisson, John Huddleston, Maika Malig, Bradley J. Nelson, Nicholas F. Parrish, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Yujun Zhang, Hugo Lam, Cristina Sisu, Danny Challis, Uday S. Evani, James Lu, Uma Nagaswamy, Jin Yu, Wangshen Li, Lukas Habegger, Haiyuan Yu, Fiona Cunningham, Ian Dunham, Kasper Lage, Jakob Berg Jespersen, Heiko Horn, Donghoon Kim, Rob Desalle, Apurva Narechania, Melissa A.Wilson Sayres, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, David Mittelman, Ruby Banerjee, Maria Cerezo, Thomas W. Fitzgerald, Sandra Louzada, Andrea Massaia, Fengtang Yang, Divya Kalra, Walker Hale, Xu Dan, Kathleen C. Barnes, Christine Beiswanger, Hongyu Cai, Hongzhi Cao, Brenna Henn, Danielle Jones, Jane S. Kaye, Alastair Kent, Angeliki Kerasidou, Rasika Mathias, Pilar N. Ossorio, Michael Parker, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Zhongming Tian, Sarah Tishkoff, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Xiaoyan Deng, Danny Asogun, Onikepe Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Tran Tinh Hien, Sarah J. Dunstan, Nguyen ThuyHang, Richard Fonnie, Robert Garry, Lansana Kanneh, Lina Moses, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Audrey Duncanson, Michael Dunn, and Jeffery A. Schloss. A global reference for human genetic variation. *Nature*, 526(7571): 68–74, 2015. ISSN 14764687. doi: 10.1038/nature15393.

[108] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomile-

hto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, Daniel G. MacArthur, H. E. Abboud, G. Abecasis, C. A. Aguilar-Salinas, O. Arellano-Campos, G. Atzmon, I. Aukrust, C. L. Barr, G. I. Bell, S. Bergen, L. Bjørkhaug, J. Blangero, D. W. Bowden, C. L. Budman, N. P. Burtt, F. Centeno-Cruz, J. C. Chambers, K. Chambert, R. Clarke, R. Collins, G. Coppola, E. J. Córdova, M. L. Cortes, N. J. Cox, R. Duggirala, M. Farrall, J. C. Fernandez-Lopez, P. Fontanillas, T. M. Frayling, N. B. Freimer, C. Fuchsberger, H. García-Ortiz, A. Goel, M. J. Gómez-Vázquez, M. E. González-Villalpando, C. González-Villalpando, M. A. Grados, L. Groop, C. A. Haiman, C. L. Hanis, A. T. Hattersley, B. E. Henderson, J. C. Hopewell, A. Huerta-Chagoya, S. Islas-Andrade, S. B. Jacobs, S. Jalilzadeh, C. P. Jenkinson, J. Moran, S. Jiménez-Morale, A. Kähler, R. A. King, G. Kirov, J. S. Kooner, T. Kyriakou, J. Y. Lee, D. M. Lehman, G. Lyon, W. MacMahon, P. K. Magnusson, A. Mahajan, J. Marrugat, A. Martínez-Hernández, C. A. Mathews, G. McVean, J. B. Meigs, T. Meitinger, E. Mendoza-Caamal, J. M. Mercader, K. L. Mohlke, H. Moreno-Macías, A. P. Morris, L. A. Najmi, P. R. Njølstad, M. C. O'Donovan, M. L. Ordóñez-Sánchez, M. J. Owen, T. Park, D. L. Pauls, D. Posthuma, C. Revilla-Monsalve, L. Riba, S. Ripke, R. Rodríguez-Guillén, M. Rodríguez-Torres, P. Sandor, M. Seielstad, R. Sladek, X. Soberón, T. D. Spector, S. E. Tai, T. M. Teslovich, G. Walford, L. R. Wilkens, and A. L. Williams. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016. ISSN 14764687. doi: 10.1038/nature19057.

[109] Philipp Rentzsch, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019. ISSN 13624962. doi: 10.1093/nar/gky1016.

[110] W. James Kent. BLAT —The BLAST -Like Alignment Tool . *Genome Research*, 12(4): 656–664, 2002. ISSN 1088-9051. doi: 10.1101/gr.229202.

[111] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 2019. ISSN 13624962. doi: 10.1093/nar/gkz369.

[112] David Zhang, Regina H Reynolds, Sonia Garcia-Ruiz, Emil K Gustavsson, Sid Sethi, Sara Aguti, Ines A Barbosa, Jack J Collier, Henry Houlden, Robert McFarland, Francesco Muntoni, Monika Oláhová, Joanna Poulton, Michael Simpson, Robert D S Pitceathly, Robert W Taylor, Haiyan Zhou, Charu Deshpande, Juan A Botia, Leonardo Collado-Torres, and Mina Ryten. Detection of pathogenic splicing events from RNA-sequencing data using dasper. *bioRxiv*, page 2021.03.29.437534, 2021. URL http://biorxiv.org/content/early/2021/03/30/2021.03.29.437534.abstract.

[113] Patricia E. Fitzsimons, Charlotte L. Alston, Penelope E. Bonnen, Joanne Hughes, Ellen Crushell, Michael T. Geraghty, Martine Tetreault, Peter O'Reilly, Eilish Twomey, Yusra Sheikh, Richard Walsh, Hans R. Waterham, Sacha Ferdinandusse, Ronald J.A. Wanders, Robert W. Taylor, James J. Pitt, and Philip D. Mayne. Clinical, biochemical, and genetic features of four patients with short-chain enoyl-CoA hydratase (ECHS1) deficiency. *American Journal of Medical Genetics, Part A*, 176(5):1115–1127, 2018. ISSN 15524833. doi: 10.1002/ajmg.a.38658.

[114] Tobias B. Haack, Christopher B. Jackson, Kei Murayama, Laura S. Kremer, André Schaller, Urania Kotzaeridou, Maaike C. de Vries, Gudrun Schottmann, Saikat Santra, Boriana Büchner, Thomas Wieland, Elisabeth Graf, Peter Freisinger, Sandra Eggimann, Akira Ohtake, Yasushi Okazaki, Masakazu Kohda, Yoshihito Kishita, Yoshimi Tokuzawa, Sascha Sauer, Yasin Memari, Anja Kolb-Kokocinski, Richard Durbin, Oswald Hasselmann, Kirsten Cremer, Beate Albrecht, Dagmar Wieczorek, Hartmut Engels, Dagmar Hahn, Alexander M. Zink, Charlotte L. Alston, Robert W. Taylor, Richard J. Rodenburg, Regina Trollmann, Wolfgang Sperl, Tim M. Strom, Georg F. Hoffmann, Johannes A. Mayr, Thomas Meitinger, Ramona Bolognini, Markus Schuelke, Jean-Marc Nuoffer, Stefan Kölker, Holger Prokisch, and Thomas Klopstock. Deficiency of ECHS1 causes

mitochondrial encephalopathy with cardiac involvement. *Annals of Clinical and Translational Neurology*, 2(5):492–509, may 2015. ISSN 23289503. doi: 10.1002/acn3.189. URL https://onlinelibrary.wiley.com/doi/10.1002/acn3.189.

[115] Michael Wainberg, Babak Alipanahi, and Brendan Frey. Does conservation account for splicing patterns? *BMC Genomics*, 17(1):1–10, 2016. ISSN 14712164. doi: 10.1186/s12864-016-3121-4. URL http://dx.doi.org/10.1186/s12864-016-3121-4.

[116] Dan Sun, Zhimei Liu, Yongchu Liu, Miaojuan Wu, Fang Fang, Xianbo Deng, Zhisheng Liu, Liang Song, Kei Murayama, Chunhua Zhang, and Yuanyuan Zhu. Novel ECHS1 mutations in Leigh syndrome identified by whole-exome sequencing in five Chinese families: Case report. *BMC Medical Genetics*, 21(1):1–12, 2020. ISSN 14712350. doi: 10.1186/s12881-020-01083-1.

[117] Beryl B Cummings, Konrad J Karczewski, Jack A Kosmicki, Eleanor G Seaby, Nicholas A Watts, Moriel Singer-Berk, Jonathan M Mudge, Juha Karjalainen, Kyle F Satterstrom, Anne ODonnell-Luria, Timothy Poterba, Cotton Seed, Matthew Solomonson, Jessica Alfoldi, The Genome Aggregation Database Production Team, The Genome Aggregation Database Consortium, Mark J Daly, and Daniel G MacArthur. Transcript expression-aware annotation improves rare variant discovery and interpretation. *bioRxiv*, page 554444, 2019. URL http://biorxiv.org/lookup/doi/10.1101/554444%0Apapers3://publication/doi/10.1101/554444.

[118] Tobias B. Haack, Christopher B. Jackson, Kei Murayama, Laura S. Kremer, André Schaller, Urania Kotzaeridou, Maaike C. de Vries, Gudrun Schottmann, Saikat Santra, Boriana Büchner, Thomas Wieland, Elisabeth Graf, Peter Freisinger, Sandra Eggimann, Akira Ohtake, Yasushi Okazaki, Masakazu Kohda, Yoshihito Kishita, Yoshimi Tokuzawa, Sascha Sauer, Yasin Memari, Anja Kolb-Kokocinski, Richard Durbin, Oswald Hasselmann, Kirsten Cremer, Beate Albrecht, Dagmar Wieczorek, Hartmut Engels, Dagmar Hahn, Alexander M. Zink, Charlotte L. Alston, Robert W. Taylor, Richard J. Rodenburg, Regina Trollmann, Wolfgang Sperl, Tim M. Strom, Georg F. Hoffmann, Johannes A. Mayr, Thomas Meitinger, Ramona Bolognini, Markus Schuelke, Jean Marc Nuoffer, Stefan Kölker, Holger Prokisch, and Thomas Klopstock. Deficiency of ECHS1 causes mitochondrial encephalopathy with cardiac involvement. *Annals of Clinical and Translational Neurology*, pages 492–509, 2015. ISSN 23289503. doi: 10.1002/acn3.189.

[119] Patricia E. Fitzsimons, Charlotte L. Alston, Penelope E. Bonnen, Joanne Hughes, Ellen Crushell, Michael T. Geraghty, Martine Tetreault, Peter O'Reilly, Eilish Twomey, Yusra Sheikh, Richard Walsh, Hans R. Waterham, Sacha Ferdinandusse, Ronald J.A. Wanders, Robert W. Taylor, James J. Pitt, and Philip D. Mayne. Clinical, biochemical, and genetic features of four patients with short-chain enoyl-CoA hydratase (ECHS1) deficiency. *American Journal of Medical Genetics, Part A*, 176(5):1115–1127, 2018. ISSN 15524833. doi: 10.1002/ajmg.a.38658.

[120] Masaaki Hayashi, Kyoko Imanaka-Yoshida, Toshimichi Yoshida, Malcolm Wood, Colleen Fearns, Revati J. Tatake, and Jiing Dwan Lee. A crucial role of mitochondrial Hsp40 in preventing dilated cardiomyopathy. *Nature Medicine*, 12(1):128–132, 2006. ISSN 10788956. doi: 10.1038/nm1327.

[121] Julian Gough, Kevin Karplus, Richard Hughey, and Cyrus Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4):903–919, 2001. ISSN 00222836. doi: 10.1006/jmbi.2001.5080.

[122] Arun Prasad Pandurangan, Jonathan Stahlhacke, Matt E. Oates, Ben Smithers, and Julian Gough. The SUPERFAMILY 2.0 database: A significant proteome update and a new webserver. *Nucleic Acids Research*, 47(D1):D490–D494, 2019. ISSN 13624962. doi: 10.1093/nar/gky1130.

[123] Shuwen Deng, Jia Liu, Xiaomei Wu, and Wei Lu. Golgi Apparatus: A Potential Therapeutic Target for Autophagy-Associated Neurological Diseases. *Frontiers in Cell and Developmental Biology*, 8(September):1–14, 2020. ISSN 2296634X. doi: 10.3389/fcell.2020.564975.

[124] Madhavi Latha Somaraju Chalasani, Asha Kumari, Vegesna Radha, and Ghanshyam Swarup. E50K-OPTN-induced retinal cell death involves the Rab GTPase-activating protein, TBC1D17 mediated block in autophagy. *PLoS ONE*, 9(4), 2014. ISSN 19326203. doi: 10.1371/journal.pone.0095758.

[125] Edward F. Griffin, Xiaohui Yan, Kim A. Caldwell, and Guy A. Caldwell. Distinct functional roles of Vps41-mediated neuroprotection in Alzheimer's and Parkinson's disease models of neurodegeneration. *Human molecular genetics*, 27(24):4176–4193, 2018. ISSN 14602083. doi: 10.1093/hmg/ddy308.

[126] Ozgur Sancak, Mark Nellist, Miriam Goedbloed, Peter Elfferich, Cokkie Wouters, Anneke Maat-Kievit, Bernard Zonnenberg, Senno Verhoef, Dicky Halley, and Ans van den Ouweland. Mutational analysis of the TSC1 and TSC2 genes in a diagnostic setting: Genotype-phenotype correlations and comparison of diagnostic DNA techniques in tuberous sclerosis complex. *European Journal of Human Genetics*, 13(6):731–741, 2005. ISSN 10184813. doi: 10.1038/sj.ejhg.5201402.

[127] Kit Sing Au, Aimee T. Williams, E. Steve Roach, Lori Batchelor, Steven P. Sparagana, Mauricio R. Delgado, James W. Wheless, James E. Baumgartner, Benjamin B. Roa, Carolyn M. Wilson, Teresa K. Smith-Knuppel, Min Yuen C. Cheung, Vicky H. Whittemore, Terri M. King, and Hope Northrup. Genotype/phenotype correlation in 325 individuals referred for a diagnosis of tuberous sclerosis complex in the United States. *Genetics in Medicine*, 9(2):88–100, 2007. ISSN 10983600. doi: 10.1097/GIM.0b013e31803068c7.

[128] Magdalena E. Tyburczy, Kira A. Dies, Jennifer Glass, Susana Camposano, Yvonne Chekaluk, Aaron R. Thorner, Ling Lin, Darcy Krueger, David N. Franz, Elizabeth A. Thiele, Mustafa Sahin, and David J. Kwiatkowski. Mosaic and Intronic Mutations in TSC1/TSC2 Explain the Majority of TSC Patients with No Mutation Identified by Conventional Testing. *PLoS Genetics*, 11(11):1–17, 2015. ISSN 15537404. doi: 10.1371/journal.pgen.1005637.

[129] Clévia Rosset, Cristina Brinckmann, Oliveira Netto, and Patricia Ashton-prolla. TSC1 and TSC2 gene mutations and their implications for treatment in Tuberous Sclerosis Complex : a review. 79:69–79, 2017.

[130] Christian Mertes, Ines Scheller, Vicente Yépez, Muhammed Çelik, Yingjiqiong Liang, Laura Kremer, Mirjana Gusic, Holger Prokisch, and Julien Gagneur. Detection of aberrant splicing events in RNA-seq data with FRASER. 2019. doi: 10.1101/2019.12.18.866830.

[131] David Zhang, Sebastian Guelfi, Sonia Garcia-Ruiz, Beatrice Costa, Regina H. Reynolds, Karishma D'Sa, Wenfei Liu, Thomas Courtin, Amy Peterson, Andrew E. Jaffe, John Hardy, Juan A. Botía, Leonardo Collado-Torres, and Mina Ryten. Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances*, 6(24):1–13, 2020. ISSN 23752548. doi: 10.1126/sciadv.aay8299.

[132] Steven L Salzberg. Open questions: How many genes do we have?, 2018. ISSN 1741-7007. URL http://www.ncbi.nlm.nih.gov/pubmed/30124169.

[133] Emmanuel Olagbaju, David Zhang, Sebastian Guelfi, and Siddharth Sethi. ODER: Optimising the Definition of Expressed Regions. 2021. doi: https://doi.org/doi:10.18129/B9.bioc.ODER.

[134] David Zhang and Leonardo Collado-Torres. dasper: Detecting abberant splicing events from RNA-sequencing data. 2020. URL http://www.bioconductor.org/packages/release/bioc/html/dasper.html.

[135] François Aguet, Alvaro N. Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E. Castel, Brian Jo, Silva Kasela, Sarah Kim-Hellmuth, Yanyu Liang, Meritxell Oliva, Elise D. Flynn, Princy Parsana, Laure Fresard, Eric R. Gamazon, Andrew R. Hamel, Yuan He, Farhad Hormozdiari, Pejman Mohammadi, Manuel Muñoz-Aguirre, Yo Son Park, Ashis Saha, Ayellet V. Segrè, Benjamin J. Strober, Xiaoquan Wen, Valentin Wucher, Kristin G. Ardlie, Alexis Battle, Christopher D. Brown, Nancy Cox, Sayantan Das, Emmanouil T. Dermitzakis, Barbara E. Engelhardt, Diego Garrido-Martín, Nicole R. Gay, Gad A. Getz, Roderic Guigó, Robert E. Handsaker, Paul J. Hoffman, Hae Kyung Im, Seva Kashin, Alan Kwong, Tuuli Lappalainen, Xiao Li, Daniel G. MacArthur, Stephen B. Montgomery, John M. Rouhana, Matthew Stephens, Barbara E. Stranger, Ellen Todres, Ana Viñuela, Gao Wang, Yuxin Zou, Shankara Anand, Stacey Gabriel, Aaron Graubert, Kane Hadley, Katherine H. Huang, Samuel R. Meier, Jared L. Nedzel, Duyen T. Nguyen, Brunilda Balliu, Donald F. Conrad, Daniel J. Cotter, Olivia M. DeGoede, Jonah Einson, Eleazar Eskin, Tiffany Y. Eulalio, Nicole M. Ferraro, Michael J. Gloudemans, Lei Hou, Manolis Kellis, Xin Li, Serghei Mangul, Daniel C. Nachun, Andrew B. Nobel, Yongjin Park, Abhiram S. Rao, Ferran Reverter, Chiara Sabatti, Andrew D. Skol, Nicole A. Teran, Fred Wright, Pedro G. Ferreira, Gen Li, Marta Melé, Esti Yeger-Lotem, Mary E. Barcus, Debra Bradbury, Tanya Krubit, Jeffrey A. McLean, Liqun Qi, Karna Robinson, Nancy V. Roche, Anna M. Smith, Leslie Sobin, David E. Tabor, Anita Undale, Jason Bridge, Lori E. Brigham, Barbara A. Foster, Bryan M. Gillard, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Ellen Karasik, Gene Kopen, William F. Leinweber, Alisa McDonald, Michael T. Moser, Kevin Myer, Kimberley D. Ramsey, Brian Roe, Saboor Shad, Jeffrey A. Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Scott D. Jewell, Daniel C. Rohrer, Dana R. Valley, David A. Davis, Deborah C. Mash, Philip A. Branton, Laura K. Barker, Heather M. Gardiner, Maghboeba Mosavel, Laura A. Siminoff, Paul Flicek, Maximilian Haeussler, Thomas Juettemann, W. James Kent, Christopher M. Lee, Conner C. Powell, Kate R. Rosenbloom, Magali Ruffier, Dan Sheppard, Kieron Taylor, Stephen J. Trevanion, Daniel R. Zerbino, Nathan S. Abell, Joshua Akey, Lin Chen, Kathryn Demanelis, Jennifer A. Doherty, Andrew P. Feinberg, Kasper D. Hansen, Peter F. Hickey, Farzana Jasmine, Lihua Jiang, Rajinder Kaul, Muhammad G. Kibriya, Jin Billy Li, Qin Li, Shin Lin, Sandra E. Linder, Brandon L. Pierce, Lindsay F. Rizzardi, Kevin S. Smith, Michael Snyder, John Stamatoyannopoulos, Hua Tang, Meng Wang, Latarsha J. Carithers, Ping Guan, Susan E. Koester, A. Roger Little, Helen M. Moore, Concepcion R. Nierras, Abhi K. Rao, Jimmie B. Vaught, and Simona Volpi. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science*, 369(6509), 2020. ISSN 10959203. doi: 10.1126/SCIENCE.AAZ5900.

[136] Dongxiao Zhu, Nan Deng, and Changxin Bai. A Generalized dSpliceType Framework to Detect Differential Splicing and Differential Expression Events Using RNA-Seq. *IEEE Transactions on Nanobioscience*, 14(2):192–202, 2015. ISSN 15361241. doi: 10.1109/TNB.2015.2388593.

[137] Fernando Pozo, Jose Manuel Rodriguez, Jesus Vazquez, and Michael L. Tress. 1 1. 2. 3.APPRIS principal isoforms and MANE Select transcripts in clinical variant interpretation. *bioRxiv*, pages 1–25, 2021.

[138] Jose Manuel Rodriguez, Fernando Pozo, Daniel Cerdán-Vélez, Tomás Di Domenico, Jesús Vázquez, and Michael L Tress. APPRIS: selecting functionally important isoforms. *Nucleic Acids Research*, pages 1–6, nov 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1058. URL https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab1058/6424762.

[139] Marco Savarese, Per Harald Jonson, Sanna Huovinen, Lars Paulin, Petri Auvinen, Bjarne Udd, and Peter Hackman. The complexity of titin splicing pattern in human adult skeletal muscles. *Skeletal Muscle*, 8(1):1–9, 2018. ISSN 20445040. doi: 10.1186/s13395-018-0156-z.

[140] Marco Savarese, Lorenzo Maggi, Anna Vihola, Per Harald Jonson, Giorgio Tasca, Lucia Ruggiero, Luca Bello, Francesca Magri, Teresa Giugliano, Annalaura Torella, Anni

Evila, Giuseppina Di Fruscio, Olivier Vanakker, Sara Gibertini, Liliana Vercelli, Alessandra Ruggieri, Carlo Antozzi, Helena Luque, Sandra Janssens, Maria Barbara Pasanisi, Chiara Fiorillo, Monika Raimondi, Manuela Ergoli, Luisa Politano, Claudio Bruno, Anna Rubegni, Marika Pane, Filippo M. Santorelli, Carlo Minetti, Corrado Angelini, Jan De Bleecker, Maurizio Moggio, Tiziana Mongini, Giacomo Pietro Comi, Lucio Santoro, Eugenio Mercuri, Elena Pegoraro, Marina Mora, Peter Hackman, Bjarne Udd, and Vincenzo Nigro. Interpreting genetic variants in titin in patients with muscle disorders. *JAMA Neurology*, 75(5):557–565, 2018. ISSN 21686149. doi: 10.1001/jamaneurol.2017.4899.

[141] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019. ISSN 14710064. doi: 10.1038/s41576-019-0093-7. URL http://dx.doi.org/10.1038/s41576-019-0093-7.

[142] Zishuai Wang, Xikang Feng, and Shuai Cheng Li. SCDevDB: A database for insights into single-cell gene expression profiles during human developmental processes. *Frontiers in Genetics*, 10(SEP):1–8, 2019. ISSN 16648021. doi: 10.3389/fgene.2019.00903.

[143] Mandeep Singh, Ghamdan Al-Eryani, Shaun Carswell, James M. Ferguson, James Blackburn, Kirston Barton, Daniel Roden, Fabio Luciani, Tri Giang Phan, Simon Junankar, Katherine Jackson, Christopher C. Goodnow, Martin A. Smith, and Alexander Swarbrick. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nature Communications*, 10(1), 2019. ISSN 20411723. doi: 10.1038/s41467-019-11049-4. URL http://dx.doi.org/10.1038/s41467-019-11049-4.

[144] Kevin Lebrigand, Virginie Magnone, Pascal Barbry, and Rainer Waldmann. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nature Communications*, 11(1):1–8, 2020. ISSN 20411723. doi: 10.1038/s41467-020-17800-6. URL http://dx.doi.org/10.1038/s41467-020-17800-6.

[145] Harriet J.A. Teare, Megan Prictor, and Jane Kaye. Reflections on dynamic consent in biomedical research: the story so far. *European Journal of Human Genetics*, 29 (4):649–656, 2021. ISSN 14765438. doi: 10.1038/s41431-020-00771-z. URL http://dx.doi.org/10.1038/s41431-020-00771-z.

[146] Rebecca Truty, Karen Ouyang, Susan Rojahn, Sarah Garcia, Alexandre Colavin, Barbara Hamlington, Mary Freivogel, Robert L. Nussbaum, Keith Nykamp, and Swaroop Aradhya. Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. *American Journal of Human Genetics*, 108(4):696–708, 2021. ISSN 15376605. doi: 10.1016/j.ajhg.2021.03.006. URL https://doi.org/10.1016/j.ajhg.2021.03.006.

[147] Michael Schatz, Anthony A. Philippakis, Enis Afgan, Eric Banks, Vincent J. Carey, Robert J. Carroll, Alessandro Culotti, Kyle Ellrott, Jeremy Goecks, Robert L. Grossman, Ira M. Hall, Kasper D. Hansen, Jonathan Lawson, Jeffrey T. Leek, Anne O'Donnell Luria, Stephen Mosher, Martin Morgan, Anton Nekrutenko, Brian D. O'Connor, Kevin Osborn, Benedict Paten, Candace Patterson, Frederick J. Tan, Casey Overby Taylor, Jennifer Vessio, Levi Waldron, Ting Wang, Kristin Wuichet, and AnVIL Team. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). *bioRxiv*, 2021. doi: https://doi.org/10.1101/2021.04.22.436044.

[148] Kendall Powell. How a field built on data-sharing became a Tower of Babel. *Nature*, 590: 198–201, 2021.

[149] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):

405–423, may 2015. ISSN 1098-3600. doi: 10.1038/gim.2015.30. URL http://www.nature.com/articles/gim201530.

[150] Revital Bronstein, Elizabeth E. Capowski, Sudeep Mehrotra, Alex D. Jansen, Daniel Navarro-Gomez, Mathew Maher, Emily Place, Riccardo Sangermano, Kinga M. Bujakowska, David M. Gamm, and Eric A. Pierce. A combined RNA-seq and whole genome sequencing approach for identification of non-coding pathogenic variants in single families. *Human Molecular Genetics*, 29(6):967–980, 2020. ISSN 14602083. doi: 10.1093/hmg/ddaa016.

[151] Tim Hagenacker, Claudia D. Wurster, René Günther, Olivia Schreiber-Katz, Alma Osmanovic, Susanne Petri, Markus Weiler, Andreas Ziegler, Josua Kuttler, Jan C. Koch, Ilka Schneider, Gilbert Wunderlich, Natalie Schloss, Helmar C. Lehmann, Isabell Cordts, Marcus Deschauer, Paul Lingor, Christoph Kamm, Benjamin Stolte, Lena Pietruck, Andreas Totzeck, Kathrin Kizina, Christoph Mönninghoff, Otgonzul von Velsen, Claudia Ose, Heinz Reichmann, Michael Forsting, Astrid Pechmann, Janbernd Kirschner, Albert C. Ludolph, Andreas Hermann, and Christoph Kleinschnitz. Nusinersen in adults with 5q spinal muscular atrophy: a non-interventional, multicentre, observational cohort study. *The Lancet Neurology*, 19(4):317–325, 2020. ISSN 14744465. doi: 10.1016/S1474-4422(20)30037-5. URL http://dx.doi.org/10.1016/S1474-4422(20)30037-5.

[152] Feng Wang, Travis Zuroske, and Jonathan K. Watts. RNA therapeutics on the rise. *Nature reviews. Drug discovery*, 19(7):441–442, 2020. ISSN 14741784. doi: 10.1038/d41573-020-00078-0. URL http://dx.doi.org/10.1038/d41573-020-00078-0.

[153] William Tapper, Victoria Hammond, Sue Gerty, Sarah Ennis, Peter Simmonds, Andrew Collins, and Diana Eccles. The influence of genetic variation in 30 selected genes on the clinical characteristics of early onset breast cancer. *Breast Cancer Research*, 10(6):1–10, 2008. ISSN 14655411. doi: 10.1186/bcr2213.

[154] Sijia Huang, Cameron Yee, Travers Ching, Herbert Yu, and Lana X. Garmire. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. *PLoS Computational Biology*, 10(9), 2014. ISSN 15537358. doi: 10.1371/journal.pcbi.1003851.

[155] Eoghan R. Malone, Marc Oliva, Peter J.B. Sabatini, Tracy L. Stockley, and Lillian L. Siu. Molecular profiling for precision cancer therapies. *Genome Medicine*, 12(1):1–19, 2020. ISSN 1756994X. doi: 10.1186/s13073-019-0703-1.

Appendix A

Supplementary Tables

	primer_name	sequence	chr	strand	start	end
1	SNCA_PF3_ER_3_P3	TCTCCTCTTACTTTGGCACTGG	chr4	-	89771561	89771582
2	SNCA_PR3_3_ER_P3	CTTCAGGTTCGTAGTCTTGATACCC	chr4	+	89726633	89726657
3	SNCA_PF2_5_ER_P2	GTGGCTGCTGCTGAGAAAACC	chr4	+	89771473	89771494
4	SNCA_PR2_ER_5_P2	CTCCAATTCTCGCCACTTCTGG	chr4	-	89835602	89835622
5	SNCA_PF4_P4	GGCATTTCATAAGCCTCATTGTC	chr4	+	89729201	89729223
6	SNCA_PR4_P4	ATCTCCTCTTACTTTGGCACTGG	chr4	-	89771561	89771583
7	SNCA_PF5_P5	AACATCAAAGGCGCTGGTTC	chr4	+	89771433	89771452
8	SNCA_PR5_P5	GCTGAGAAGACCAAAGAGCAAG	chr4	-	89822365	89822386

Table A.1 Primer positions and sequences used to experimentally validate the novel ER of SNCA

	Ensembl Transcript ID	Protein Start	Protein End	Domain source	Description
1	ENST00000262375	80.00	457.00	PANTHER	DNAJ HOMOLOG SUBFAMI
2	ENST00000262375	80.00	457.00	PANTHER	DNAJ HOMOLOG SUBFAMI
3	ENST00000262375	207.00	416.00	CDD	DnaJ_C
4	ENST00000262375	211.00	321.00	Gene3D	Urease metallochaperone UreE
5	ENST00000262375	236.00	296.00	Gene3D	-
6	ENST00000262375	335.00	438.00	Gene3D	Urease metallochaperone UreE
7	ENST00000262375	90.00	477.00	HAMAP	DnaJ
8	ENST00000262375	210.00	413.00	Pfam	DnaJ_C
9	ENST00000262375	89.00	203.00	Gene3D	Chaperone J-domain superfam
10	ENST00000262375	90.00	196.00	SuperFamily	Chaperone J-domain
11	ENST00000262375	92.00	150.00	Smart	dnaj_3
12	ENST00000262375	93.00	147.00	CDD	DnaJ
13	ENST00000262375	93.00	155.00	Pfam	DnaJ
14	ENST00000262375	93.00	158.00	Prosite_profiles	DNAJ_2
15	ENST00000262375	95.00	113.00	PRINTS	JDOMAIN
16	ENST00000262375	113.00	128.00	PRINTS	JDOMAIN
17	ENST00000262375	130.00	150.00	PRINTS	JDOMAIN
18	ENST00000262375	150.00	169.00	PRINTS	JDOMAIN
19	ENST00000262375	135.00	154.00	Prosite_patterns	DNAJ_1
20	ENST00000262375	338.00	424.00	SuperFamily	HSP40/DnaJ peptide-binding c
21	ENST00000262375	223.00	301.00	Prosite_profiles	ZF_CR
22	ENST00000262375	236.00	296.00	CDD	DnaJ_zf
23	ENST00000262375	236.00	296.00	Pfam	DnaJ_CXXCXGXG
24	ENST00000262375	223.00	299.00	SuperFamily	DnaJ/Hsp40 cysteine-rich dom

Table A.2 Protein domains detected within the *DNAJA3* transcript ENST00000262375. This table is downloaded through the Ensembl web browser and contains all protein coding domains that were present in the *DNAJA3* MANE-select transcript ENST00000262375. Protein start and Protein end columns specify the amino acid positions where the corresponding domain starts and ends.

	Sample ID	Provider	Sex	Ethnicity	Age of onset of symptoms
1	M0367-16	Newcastle	M	Pakistani (consanguineous)	Infancy
2	M2566-15	Newcastle	M	Nepalese (consanguineous)	Birth
3	M1237-16	Newcastle	F	Not stated (consanguineous)	Birth
4	M0906-17	Newcastle	F	South African	Infancy
5	M1451-17	Newcastle	F	Not stated (consanguineous)	Infancy
6	M1532-17	Newcastle	M	Irish	Infancy
7	M1316-12	Newcastle	M	Not stated (consanguineous)	Birth
8	M2198-15	Newcastle	F	European	Birth
9	M0905-18	Newcastle	F	European	Birth
10	M0014-18	Newcastle	F	European	Infancy
11	M1708-15	Newcastle	M	European	Infancy
12	M0892-14	Newcastle	F	European	Infancy
13	M0138-11	Newcastle	M	European	Infancy
14	M0687-14	Newcastle	F	European	Infancy
15	M1122-11	Newcastle	F	Pakistani (consanguineous)	Infancy
16	M0229-16	Newcastle	M	Not stated	Birth
17	M1710-16	Newcastle	F	European	Birth
18	ION176	UCL	M	Caucasian	Adult
19	L1219.1875F	UCL	M	South Asian	Adult
20	L1550.2631F	UCL	F	Caucasian	Childhood
21	L1901.3262F	UCL	F	European	Childhood
22	L949.3246F	UCL	M	Irish (consanguineous)	Birth
23	S1741	Oxford	M	Not stated	Infancy
24	S1742	Oxford	M	Not stated	Birth
25	S1743	Oxford	M	Not stated	Infancy
26	S1820	Oxford	F	Not stated	Infancy
27	S2110	Oxford	F	Caucasian	Infancy
28	S2112	Oxford	M	Not stated	Birth
29	S2220	Oxford	M	Not stated	Birth
30	S2220B	Oxford	M	Not stated	Birth
31	S2582	Oxford	M	Not stated	Infancy
32	S2586BK	Oxford	M	Turkish	Birth

Table A.3 The sample identifiers, institutions that provided samples and the demographic of the patient samples with suspected mitochondrial disorders analysed in chapter 4.

	Sample ID	Phenotype
1	M0367-16	Encephalopathy;Liver disease;Coagulation defect;MRI brain showing basal ganglia changes;Affected brother died age 10 weeks
2	M2566-15	Hypotonia;Encephalopathy;MRI brain showing widespread cystic changes;Apnoeic
3	M1237-16	Collapse at 18hrs of age;Hyperkalemia, elevated lactate;Renal failure; arrhythmia;Sinus arrest;Recurrent profound bradycardia, endocrinopathy ;RIP
4	M0906-17	Global neurodevelopmental delay;Poor swallow;Myopathic facies;Generalised hypotonia;Recurrent admission to PICU following viral illnesses;Predominate type 1 fibres
5	M1451-17	Leigh syndrome;Global developmental delay;PEG fed;2 affected sibs also died;RIP age 12 months.
6	M1532-17	Premature baby due to IUGR;SUDI at corrected age 6/52
7	M1316-12	Down syndrome;Unexplained progressive muscle weakness;CPAP dependent
8	M2198-15	IUGR; born 34w;Seizures;Developmental delay;Failure to thrive;Bilateral ptosis;Infantile spasms;Hypotonia;RIP at 7months.
9	M0905-18	Dilated congenital cardiomyopathy;Hydropic at birth (born at 29 + 3w);RIP.
10	M0014-18	Encephalopathy;Seizures;PEG fed;Hypertrichosis;Learning difficulties;Short stature (<0.4th centile);Visual loss ;Renal impairment;Brain MRI showing cerebellar and cerebral atrophy;Marked deterioration in last 2y.
11	M1708-15	Multiple cranial neuropathies;Poor bulbar control;Reflux; hearing loss;Demyelinating sensory and motor neuropathy
12	M0892-14	Liver steatosis;Hepatomegaly
13	M0138-11	Clinical and radiological evidence of Leigh syndrome;Affected brother died age 4y;Patient RIP age 18y (Leigh Syndrome)
14	M0687-14	Developmental regression;Consistent with Leigh syndrome affecting younger sister
15	M1122-11	Developmental regression at 3 months following illness;Dysphagia and loss of head control; developmental delay;Hypotonia;Encephalopathy;Myopathy;MRI brain showing bilateral symmetrical diffuse leukodystrophy
16	M0229-16	Hypotonia;Muscle weakness since birth;Apnoeas; hiccoughs;Absent reflexes;Severe global developmental delay;Normal EEG
17	M1710-16	Hypotonia;Fixed dilated pupils;Ventilated at 9hrs of life;MRI brain showing signal changes in basal ganglia and dorsal brainstem.
18	ION176	Muscle cramps at age 27 after exertion;One possible episode of rhabdomyolysis;Fatigue;Myalgia;Unilateral ptosis;Migraine;Diet-controlled diabetes
19	L1219.1875F	Complex neurological presentation with onset in late 30s;Optic atrophy;Hearing loss;Focal seizures;Ataxia;Muscle cramps;Bladder and bowel symptoms;Small fibre neuropathy on nerve conduction studies;MRI brain showing periventricular white matter changes
20	L1550.2631F	Onset late childhood;Leukodystrophy;Severe constipation;Myopathy;Neuropathy complex iv deficiency
21	L1901.3262F	Muscle weakness;ptosis;dystonia;PEG fed;deafness;diabetes
22	L949.3246F	Hypotonia;Delayed motor skills;Dystonia; spastic paraparesis;Learning difficulties;Kyposcoliosis;MRI brain showing possible mineralisation
23	S1741	Suspected mitochondrial disorder
24	S1742	Poor nutrition with gastro-oesophageal reflux post Nissen fundoplication;Congenital myopathy with ptosis;Fatigue;Hypoglycaemia;Palpitations treated with bisoprolol
25	S1743	Suspected mitochondrial disorder
26	S1820	Severe infantile-onset dystonia and dyskinesias with choreoathetosis;Microcephaly;Possible sensory neuropathy;Bulbar dysfunction; epilepsy;Possible pre-excitation with SVTs;Subtle thalamic changes on MRI
27	S2110	Motor delay;Ptosis;Fatigue requiring wheelchair
28	S2112	Congenital myopathy;Bilateral ptosis;External opthalmoplegia;Hypotonia;Fatigue and poor exercise tolerance;PEG fed;Stiff spine and rotational scoliosis;Poor voice quality and malocclusion
29	S2220	Congenital haemochromatosis;RIP at < 1yr
30	S2220B	Congenital haemochromatosis;RIP at < 1yr
31	S2582	Severe motor neuronopathy with anterior horn cell features;PEG fed from 2 years;Respiratory failure with bronchiectasis;Profound intellectual disability;Severe epilepsy;Thin corpus callosum;Osteoporosis with multiple fractures
32	S2586BK	Hypotonia; poor feeding; epilepsy;Coarse facial features;Hypertonia;Mild contractures of knees and ankles;Abnormal breathing pattern

Table A.4 The clinical phenotype of the samples with suspected mitochondrial disorders analysed in chapter 4.

	Sample ID	Biochemistry
1	M0367-16	Complex I, III, IV deficiency in skeletal muscle Elevated blood lactate
2	M2566-15	Normal respiratory chain enzymes in fibroblasts Elevated lactate Normal acylcarnitines
3	M1237-16	Deficiency of myristate Palmitate Oleate Marked mtDNA depletion in skeletal muscle
4	M0906-17	Normal respiratory chain enzymes in skeletal muscle Predominate type 1 fibres
5	M1451-17	Normal respiratory chain enzymes in skeletal muscle Elevated blood and CSF lactate Normal PDH in muscle
6	M1532-17	Complex I deficiency in skeletal muscle mtDNA copy number normal
7	M1316-12	Complex I and IV deficiency in skeletal muscle COX deficient fibres
8	M2198-15	Complex I deficiency in fibroblasts Elevated blood lactate and CSF lactate
9	M0905-18	Complex I deficiency in muscle
10	M0014-18	Complex I deficiency in skeletal muscle Elevated blood and CSF lactate Previously glycine receptor antibody positive Previously low IGF1 Increased alanine
11	M1708-15	Complex I deficiency in skeletal muscle
12	M0892-14	Mild complex I deficiency in skeletal muscle Elevated plasma lactate Elevated LFTs
13	M0138-11	Normal respiratory chain enzymes in muscle and fibroblasts Normal PDH mtDNA depletion excluded
14	M0687-14	Complex I deficiency in muscle (normal in fibroblasts)
15	M1122-11	Complex I deficiency in muscle Normal respiratory chain enzymes in fibroblasts
16	M0229-16	Complex IV deficiency Normal CSF lactate Normal creatinine kinase
17	M1710-16	Mild complex I deficiency in skeletal muscle Normal blood lactate Normal CSF lacate
18	ION176	Complex IV deficiency COX negative fibres
19	L1219.1875F	Respiratory chain enzymes not checked; multiple mtDNA deletions; muscle histology normal
20	L1550.2631F	Complex IV deficiency
21	L1901.3262F	Normal respiratory chain enzymes; elevated lactate COX negative fibres Non-specific myopathic changes Previous detection of mitochondrial DNA deletions
22	L949.3246F	mtDNA deletions COX negative fibres
23	S1741	-
24	S1742	Respiratory chain enzyme deficiency in skeletal muscle
25	S1743	-
26	S1820	Complex IV deficiency with borderline low complex I deficiency No evidence of mtDNA depletion
27	S2110	Complex IV deficiency Borderline raised blood lactate
28	S2112	Respiratory chain enzyme deficiency COX negative fibres
29	S2220	Low mtDNA copy number in skeletal muscle
30	S2220B	Low mtDNA copy number in skeletal muscle
31	S2582	-
32	S2586BK	Respiratory chain enzyme deficiency in skeletal muscle CSF neurotransmitters low

Table A.5 The biochemistry results of the samples with suspected mitochondrial disorders analysed in chapter 4.

	Sample ID	Diagnostic selection
1	M0367-16	WES -ve; mtDNA normal
2	M2566-15	WES -ve
3	M1237-16	Diagnosis established after cell line submission: homozygous variant in SLC25A20 (c.713A>G p.Gln238Arg)
4	M0906-17	WES -ve; mtDNA normal
5	M1451-17	WES -ve; mtDNA normal
6	M1532-17	WES -ve
7	M1316-12	WES -ve
8	M2198-15	Trio WES: de novo heterozygous variant in NDUFS5 and DNM1L classified as VUSs
9	M0905-18	WES: single heterozygous, pathogenic variant in NDUFS8
10	M0014-18	Complex I panel -ve; WES -ve
11	M1708-15	Complex I panel: single heterozygous variant NDUFAF1 variant classified as VUS; WES -ve
12	M0892-14	WES -ve; Normal mtDNA copy number
13	M0138-11	m.14465G>A MTND6; both brothers homoplasmic; sibling with autism spectrum disorder and mother heteroplasmic
14	M0687-14	WES: compound heterozygous variants in NDUFS7 classified as VUS (c.256C>G p.Leu86Val and c.334G>A p.Ala112Thr)
15	M1122-11	Complex I panel: homozygous 12bp duplication NDUFV2 variant classified as VUS
16	M0229-16	WES: compound heterozygous variants in SQRDL
17	M1710-16	WES: heterozygous variant in NDUFS3 classified as a VUS
18	ION176	WGS -ve
19	L1219.1875F	WGS -ve
20	L1550.2631F	WGS -ve
21	L1901.3262F	WGS -ve
22	L949.3246F	WGS; MIEF1 candidate
23	S1741	WES -ve; mtDNA normal
24	S1742	WES -ve; mtDNA normal
25	S1743	WES -ve; mtDNA normal
26	S1820	WES -ve; mtDNA normal
27	S2110	WES -ve; mtDNA normal
28	S2112	WES -ve; mtDNA normal
29	S2220	WES -ve; mtDNA normal
30	S2220B	WES -ve; mtDNA normal
31	S2582	WES -ve; mtDNA normal
32	S2586BK	WES -ve; mtDNA normal

Table A.6 The diagnostic criteria for the samples with suspected mitochondrial disorders analysed in chapter 4. This includes the details of any diagnoses or discovered VUSs during the thesis.