Fakultät für Informatik
Technische Universität München

Improving the genetic diagnosis of Mendelian disorders via robust outlier detection methods for transcriptome sequencing

Christian Mertes

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Burkhard Rost

Prüfende der Dissertation:
1. Prof. Dr. Julien Gagneur
2. Prof. Dr. Thomas Meitinger

Die Dissertation wurde am 05.07.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 13.10.2021 angenommen.

Acknowledgments

My deepest thanks and appreciation goes to Prof. Julien Gagneur. Starting with my Master's thesis in your lab, you have always been there as a mentor on a scientific level, but also personally. Thank you for trusting me and challenging me to bring me to the next level. Over the years, you have managed to build a vibrant, fun, but also a scientifically challenging lab that I am very grateful to have been a part of, as the years have produced a lot of great memories even though I was away for half of my PhD. Without Daniel Bader's persuasion to apply for a master's thesis in the Gagneur lab, I would not have ended up where I am now. So thank you for introducing me to the Gagneur lab! You were always there as a friend, but also as a colleague during my years in Munich. I would also like to thank Vicente in particular, but also Leo, Jun, Juri and Žiga. You all are the colleagues one would wish to have. I have so many great memories with all of you, from the party boat and hiking in the mountains, to *someone* is explaining the world, to all the paper parties. But also scientifically, you helped answering all my questions, collaborated on cool projects, kept the bar high, and pushed new ideas. A big thank you also goes to Felix and Ines, without your tireless work and help in optimizing outlier detection in RNA-seq and the countless discussions I probably wouldn't be here now. Even though I supervised your master thesis, you taught me a lot. Thank you Flo for taking over the task of the system administrator so passionately. I would also like to thank Inga who solved all (my) administrative problems. All of you, including the new lab members, make this place special. And thanks to all of you who put up with me when some or all of the servers were down ;)

Further, I thank all my collaborators, especially the Prokisch lab. Thank you Dr. Holger Prokisch for the interesting and critical scientific discussions, which always brought up new ideas, but also led to better science. I would like to thank Laura, Mirjana, Sarah, Robert, Riccardo and all the others for their collaboration. Your biological knowledge and your critical reviews and manual investigations of all our results are much appreciated. By the way, without you the retreats would have been half the fun!

Furthermore, I would like to thank Prof. Lars Steinmetz and Prof. Stephen Montgomery for supporting me and making my dream of working abroad come true, but also Prof. Julien Gagneur for letting me go for so long in the first place. It was a great pleasure to work in your labs and to collaborate with Laure Frésard on the RNA-seq workshop.

Finally, I am grateful to my family and friends who have supported me all these years and made sure that the work-life balance was right. Also, many thanks to my old teachers who sparked in me an interest in biology and mathematics. And finally, I thank you, Anne, from the bottom of my heart, as you always stood behind me and helped and pushed me wherever needed.

Summary

Finding and understanding the pathomechanism in rare genetic diseases is important for diagnosis, genetic counseling, treatment, and development of therapies. This requires pinpointing the underlying genetic cause and its downstream effects. Over the last decades, whole-exome sequencing and whole-genome sequencing revolutionized the diagnostic field by increasing the diagnostic yield, shortening the turnover time, and accelerating the discovery of novel disease genes. Nonetheless, this still leaves most individuals with a suspected Mendelian disease undiagnosed due to our current limitations in predicting and interpreting the effect for most of the variants an individual harbors. In this thesis, I will address these limitations by presenting how RNA sequencing as a complementary avenue can be used to improve further the diagnostic yield by directly probing the functional effects of genetic variation. Further, I will present robust statistical algorithms to detect aberrant gene expression and splicing events to support RNA sequencing-based diagnostics.

First, I systematically investigated in a pilot study together with colleagues how RNA sequencing can be used to detect aberrant gene expression, aberrant splicing events, and mono-allelic expression of a rare variant. To this end, I adapted algorithms developed for differential gene expression and splicing analysis to the rare disease scenario. I applied these algorithms to RNA sequencing data of 105 individuals to detect aberrant events. This led to the diagnosis in 10% of previously undiagnosed cases by revealing the underlying genetic cause and pathomechanism.

In a diagnostic setting, it is important to have robust, reproducible, and trustworthy predictions. Therefore, I developed OUTRIDER and FRASER, specialized algorithms to detect aberrant gene expression and splicing events in RNA sequencing data. I implemented denoising autoencoders to control for known and unknown confounders and used appropriate distributions to assess the significance of aberrant events. I extensively tested, compared, and benchmarked the developed methods using simulated and experimental datasets.

Altogether, with this work, I have demonstrated the benefits and complementarity of RNA sequencing-based diagnostics and developed robust algorithms to detect aberrant events in RNA sequencing data. With this, I have directly contributed to the diagnosis of more than 40 individuals over the past several years. As RNA sequencing-based diagnostics are rapidly adopted by the rare disease community, my open-source packages are being used worldwide to improve the diagnostic rate. Complementary RNA sequencing-based diagnostics ultimately improve the quality of life of patients through molecular diagnosis and thus genetic counseling, treatment, and development of therapies.

Publications

Major contributions:

Genetic diagnosis of Mendelian disorders via RNA sequencing

Ref. Kremer et al. [2017]

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, Agnès Rötig, Peter Freisinger, Felix Distelmaier, Tim M. Strom, Thomas Meitinger, Julien Gagneur, and Holger Prokisch

(2017) Nature Communications, DOI: 10.1038/ncomms15824.

Author contribution T.M., J.G. and H.P. planned the project. J.G. and H.P. overviewed the research. H.P. designed the experiments. C.L., B.F., A.D., V.T., A.L., D.G., R.W.T., D.G., J.A.M., A.R., P.F., F.D. and T.M. reviewed the phenotypes, performed sample collection and biochemical analysis. L.S.K., D.M.B., C.M., T.M.S. and H.P. curated and analysed the data. J.G. devised the statistical analysis. L.S.K., R.K., A.I., C.T., E.K. and B.R. performed the cell biology experiments. L.S.K., R.K., E.G., T.S., P.L. and T.M.S. performed exome, genome and RNA-seq. L.S.K., R.K., T.B.H. and H.P. performed the exome analysis. L.S.K. and G.P. performed the quantitative proteomics experiments. L.S.K., G.K. and J.A. performed the metabolomics studies. L.S.K., D.M.B., C.M., J.G. and H.P. wrote the manuscript. L.S.K., D.M.B. and C.M. visualized the data. Critical revision of the manuscript was performed by all authors.

OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data

Ref. Brechtmann et al. [2018]

Felix Brechtmann,* **Christian Mertes**,* Agnė Matusevičiūtė*, Vicente A. Yépez, Žiga Avsec, Maximilian Herzog, Daniel M. Bader, Holger Prokisch, and Julien Gagneur.

(2018) American Journal of Human Genetics, DOI: 10.1016/j.ajhg.2018.10.025.

Author contribution J.G. conceived the project and overviewed the research with the help of Ž.A., H.P., and V.A.Y.. F.B., A.M., C.M. analyzed the data. F.B., C.M. and A.M. developed the software. D.M.B. M.H contributed to the software development and early stage data analysis. J.G. and Ž.A. devised the statistical analysis. F.B., V.A.Y., C.M., and J.G. made the figures. F.B., C.M., A.M., V.A.Y. and J.G. wrote the manuscript. All authors performed critical revision of the manuscript.

Detection of aberrant splicing events in RNA-seq data with FRASER

Ref. Mertes et al. [2021]

Christian Mertes,* Ines F. Scheller,* Vicente A. Yépez, Muhammed H. Çelik, Yingjiqiong Liang, Laura S. Kremer, Mirjana Gusic, Holger Prokisch, and Julien Gagneur

(2021) Nature Communications, DOI: 10.1038/s41467-020-20573-7.

Author contribution C.M. and J.G. conceived the method. C.M. and I.S. implemented the package and performed the full analysis. V.A.Y. contributed to the package development and to the analysis. M.H.Ç. performed the MMSplice analysis of GTEx. C.M. and Y.L. performed the rare variant enrichment analysis. L.S.K. and M.G. analyzed the results of the rare disease cohort. J.G and H.P. supervised the research. C.M., I.S, and J.G. wrote the manuscript with the help of V.A.Y. All authors revised the manuscript.

Minor contributions:

Chromatin-remodeling factor SMARCD2 regulates transcriptional networks controlling differentiation of neutrophil granulocytes

Ref. Witzel et al. [2017]

Maximilian Witzel, Daniel Petersheim, Yanxin Fan, Ehsan Bahrami, Tomas Racek, Meino Rohlfs, Jacek Puchałka, **Christian Mertes**, Julien Gagneur, Christoph Ziegenhain, Wolfgang Enard, Asbjørg Stray-Pedersen, Peter D. Arkwright, Miguel R. Abboud, Vahid Pazhakh, Graham J. Lieschke, Peter M. Krawitz, Maik Dahlhoff, Marlon R. Schneider, Eckhard Wolf, Hans-Peter Horny, Heinrich Schmidt, Alejandro A. Schäffer, and Christoph Klein

(2017) Nature Genetics, DOI: 10.1038/ng.3833.

Author contribution M.W. designed, performed, and interpreted experiments and wrote and edited the manuscript. D.P. performed ATAC–seq and RNA–seq, Y.F., E.B., T.R., and M.R. were involved in genomic and biochemical analyses, J.P. led the computational biology efforts, C.M. and J.G. analyzed ATAC–seq and RNA–seq data, and C.Z. and W.E. performed mouse RNA–seq and digital gene expression analysis. A.S.-P., P.D.A., and M.R.A. provided clinical care for patients, V.P. and G.J.L. generated and analyzed zebrafish models, and P.M.K. analyzed whole-exome sequencing in initial patients. M.D., M.R.S., and E.W. generated mice. H.-P.H. performed immunohistochemistry analysis of bone marrow biopsies, H.S. provided expert clinical genetic consulting, and A.A.S. guided bioinformatics studies and helped write and edit the manuscript. C.K. designed and guided the study, supervised M.W., provided laboratory resources, and wrote the manuscript.

Mutations in *MDH2*, Encoding a Krebs Cycle Enzyme, Cause Early-Onset Severe Encephalopathy

Ref. Ait-El-Mkadem et al. [2017]

Samira Ait-El-Mkadem, Manal Dayem-Quere, Mirjana Gusic, Annabelle Chaussenot, Sylvie Bannwarth, Bérengère François, Emmanuelle C. Genin, Konstantina Fragaki, Catharina L.M. Volker-Touw, Christelle Vasnier, Valérie Serre, Koen L.I. van Gassen, Françoise Lespinasse, Susan Richter, Graeme Eisenhofer, Cécile Rouzier, Fanny Mochel, Anne De Saint-Martin, Marie-Thérèse Abi Warde, Monique G.M. de Sain-van der Velde, Judith J.M. Jans, Jeanne Amiel, Žiga Avsec, **Christian Mertes**, Tobias B. Haack, Tim

Strom, Thomas Meitinger, Penelope E. Bonnen, Robert W. Taylor, Julien Gagneur, Peter M. van Hasselt, Agnès Rötig, Agnès Delahodde, Holger Prokisch, Sabine A. Fuchs, and Véronique Paquis-Flucklinger

(2017) American Journal of Human Genetics, DOI: 10.1016/j.ajhg.2016.11.014.

Author contribution Ž.A. and C.M. developed and implemented GenePROF. C.M. and M.G. analyzed results of GenePROF. All authors revised the manuscript.

Somatic alterations compromised molecular diagnosis of DOCK8 hyper-IgE syndrome caused by a novel intronic splice site mutation

Ref. Hagl et al. [2018]

Beate Hagl, Benedikt D. Spielberger, Silvia Thoene, Sophie Bonnal, **Christian Mertes**, Christof Winter, Isaac J. Nijman, Shira Verduin, Andreas C. Eberherr, Anne Puel, Detlev Schindler, Jürgen Ruland, Thomas Meitinger, Julien Gagneur, Jordan S. Orange, Marielle E. van Gijn, and Ellen D. Renner

(2018) Scientific Reports, DOI: 10.1038/s41598-018-34953-z.

Author contribution B.H., B.D.S., S.T., S.B., A.C.E., A.P., D.S., performed research and analyzed data; C.M., C.W., I.J.N., S.V., J.R., T.M., J.G., J.S.O., M.E.v.G. analyzed data; B.H., B.D.S., E.D.R. analyzed clinical data; B.H., B.D.S. performed STAT3 analysis and post-sort gDNA sequencing; B.H. performed cDNA analysis, minigene analysis and isolated fibroblasts, B.D.S. performed T subpopulation, DOCK8 protein expression and autoantibody analysis; B.H., B.D.S., E.D.R. designed the research and were the principal writers of the manuscript. All of the authors reviewed the manuscript and contributed in writing.

Detection of aberrant gene expression events in RNA sequencing data

Ref. Yépez et al. [2021b]

Vicente A. Yépez, **Christian Mertes**, Michaela F. Müller, Daniela S. Andrade, Leonhard Wachutka, Laure Frésard, Mirjana Gusic, Ines F. Scheller, Patricia F. Goldberg, Holger Prokisch, and Julien Gagneur

(2020) Nature Protocols, DOI: 10.1038/s41596-020-00462-5

Author contribution Participated in the design of the workflow: V.A.Y., C.M., M.F.M., and J.G.. Contributed to the computational workflow: V.A.Y., C.M., M.F.M., D.S.A., I.S., and P.F.G.. Implemented the candidate prioritization workflow: L.F.. Designed and implemented wBuild: L.W.. Wrote the manuscript: V.A.Y. and J.G.. All authors revised the manuscript.

Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders

Ref. Kopajtich et al. [2021]

Robert Kopajtich,* Dmitrii Smirnov,* Sarah L. Stenton,* Stefan Loipfinger, Chen Meng, Ines F. Scheller, Peter Freisinger, Robert Baski, Riccardo Berutti, Jürgen Behr, Martina Bucher, Felix Distelmaier, Mirjana Gusic, Maja Hempel, Lea Kulterer, Johannes Mayr, Thomas Meitinger, **Christian Mertes**, Metodi D. Metodiev, Agnieszka Nadel, Alessia Nasca, Akira Ohtake, Yasushi Okazaki, Rikke Olsen, Dorota Piekutowska-Abramczuk, Agnès Rötig, René Santer, Detlev Schindler, Abdelhamid Slama, Christian Staufner, Tim Strom, Patrick Verloo, Jürgen-Christoph von Kleist-Retzow, Saskia B. Wortmann, Vicente A. Yépez, Costanza Lamperti, Daniele Ghezzi, Kei Murayama, Christina Ludwig, Julien Gagneur, and Holger Prokisch

(2021) medRχiv, DOI: 10.1101/2021.03.09.21253187

Author contribution Conceived and supervised the study, H.P; performed experiments, R.K, L.K, C.Lu, D.G, and M.M; analyzed and interpreted results, D.S, S.L, I.S, C.M, V.Y, D.G, M.M, R.K, S.L.S, J.G, H.P; provided essential materials, all authors; wrote the manuscript, S.L.S, H.P, J.G, R.K, and D.S; edited manuscript, all authors.

Clinical implementation of RNA sequencing for Mendelian disease diagnostics

Ref. Yépez et al. [2021a]

Vicente A. Yépez,* Mirjana Gusic,* Robert Kopajtich, **Christian Mertes**, Nicholas H. Smith, Charlotte L. Alston, Riccardo Berutti, Holger Blessing, Elżbieta Ciara, Fang Fang, Peter Freisinger, Daniele Ghezzi, Susan J. Hayflick, Yoshihito Kishita, Thomas Klopstock, Costanza Lamperti, Dominic Lenz, Christine C. Makowski, Johannes A. Mayr, Signe Mosegaard, Michaela F. Müller, Gerard Muñoz-Pujol, Kei Murayama, Agnieszka Nadel, Akira Ohtake, Yasushi Okazaki, Dorota Piekutowska-Abramczuk, Elena Procopio, Antonia Ribes, Agnès Rötig, Thomas Schwarzmayr, Christian Staufner, Sarah L. Stenton, Tim M. Strom, Robert W. Taylor, Caterina Terrile, Frederic Tort, Rudy Van Coster, Matias Wagner, Saskia B. Wortmann, Manting Xu, Thomas Meitinger, Julien Gagneur, and Holger Prokisch

(2021) medRχiv, DOI: 10.1101/2021.04.01.21254633

Author contribution Conceptualization: JG, HP. Data Curation Management: VAY, MG, RK, AN. Formal Analysis: VAY, MG, RK, CM. Investigation: MG, RK, AN. Resources: CLA, HB, EC, FF, PF, DG, SJH, YK, TK, CL, DL, CCM, JAM, SM, GMP, KM, AO, YO, DPA, EP, ARi, ARo, CS, RWT, CT, FT, RVC, MW, SW, MX. Software: VAY, CM, NHS, MFM, RB, TS. Supervision: JG, HP. Validation: MG, RK, SLS, HP. Visualization: VAY, MG, JG, HP. Writing – Original Draft Preparation: VAY, MG, JG, HP. Writing – Review and Editing: all authors.

Contents

1 Introduction

> *Nature is nowhere accustomed more openly to display her secret mysteries than in cases where she shows tracings of her workings apart from the beaten paths; nor is there any better way to advance the proper practice of medicine than to give our minds to the discovery of the usual law of nature, by careful investigation of cases of rarer forms of disease.*
>
> William Harvey, 1657

It is estimated that worldwide between 263-446 million people are living with one of the 6000-8000 defined rare diseases.[Nguengang Wakap et al., 2020] This means between 3.5-5.9% of the world's population is living with a disease that is severe and disabling (66%), life-limiting (50%), and in most cases untreatable or curable (94%).[Boycott and Ardigó, 2018] Children are the most vulnerable demographic population as every second affected individual is a child. Hence, multiple national and international efforts are raising awareness, facilitating rare disease research and drug development, and making rare diseases a public health priority (US' Orphan Drug Act, 1983; EU's Orphan Medicinal Products, 1999;[1] EURORDIS, 2005;[Rode, 2005] IRDiRC, 2011[Boycott et al., 2017]). As more than 50% of the individuals living with a rare disease do not receive a molecular diagnosis,[Neveling et al., 2013; Wortmann et al., 2015; Wright et al., 2018b; Retterer et al., 2016] the IRDiRC has the ambitious goal to provide a diagnosis to any individual with a known rare disease within a year.[Boycott et al., 2017] To achieve this, new technologies and strategies have to be developed and implemented, which is the aim of my thesis through advancing RNA sequencing (RNA-seq)-based diagnostics.

1.1 Rare diseases and rare genetic disorders

Rare diseases are rare in themselves, but together they are common with a prevalence of 3.5-5.9%.[Nguengang Wakap et al., 2020] Despite the estimated 263-446 million people living with a rare disease worldwide, no common definition for it exists.[Richter et al., 2015] The EU defines a rare disease as life-threatening or chronically debilitating conditions where less than 1 in 2,000 people are affected.[1] In the US, it is defined in absolute

[1] Regulation (EC) No 141/2000 of the European parliament and of the council of 16 December 1999 on orphan medicinal products. `http://ec.europa.eu/health//sites/health/files/files/eudralex/vol-1/reg_2000_141_cons-2009-07/reg_2000_141_cons-2009-07_en.pdf`

terms with fewer than 200,000 affected people translating to 1 in $\sim 1,600$.[2] We do not know the full spectrum of rare diseases yet, but till now more than 6000 known rare diseases are defined and the number is growing each year.[Amberger et al., 2019][3] It is estimated that 69.9% of them are exclusively pediatric onset, while 71.9% are genetic, i.e. caused by alterations in the genome that lead to harmful changes in the function of single genes.[Nguengang Wakap et al., 2020] The latter subgroup of rare diseases is referred to as rare genetic disorders, Mendelian diseases, or monogenic diseases. Since this thesis focuses on rare genetic diseases, I will use the term *rare diseases* interchangeably for this subset in the following for simplicity. Being affected by or diagnosed with a rare disease can be devastating, since most of them have an enormous negative impact on the well-being of the affected person himself, but also on those around him. But the reality of rare diseases is even worse with two-thirds being disabling, three-quarters affecting children, over half being life-limiting, and most without a treatment.[Boycott and Ardigó, 2018] In addition, it is estimated that in half of the rare genetic diseases, the underlying etiology has yet to be discovered.[Boycott et al., 2017] A timely molecular diagnosis is important in many ways. The early understanding of the disease can improve the disease management through targeted therapies and hence reduce or delay long-term complications. Moreover, it is essential to know the underlying mechanism in order to develop targeted drugs. Currently, only 6% of the rare diseases have approved treatments.[Austin et al., 2018] A proper molecular diagnosis also reduces prognostic uncertainty and provides better means for genetic counseling. Thus, one of IRDiRC's ambitious goals is to provide a diagnosis for all individuals with a known rare genetic disease, as the rate of diagnosis is directly related to the successful implementation of precision medicine.[Boycott et al., 2019]

1.2 Finding the needle in the haystack: molecular diagnosis in rare diseases

Molecular diagnostics is a broad term for analyzing biomarkers in human samples in a clinical setting.[Poste, 2014] Some of the most used techniques in molecular diagnostics are next-generation sequencing (NGS), mass spectrometry, enzyme-linked immunosorbent assays (ELISA), and fluorescence in situ hybridization (FISH). In the context of rare genetic diseases, the molecular diagnosis is often referred to as the identification of the causal genetic defect on the DNA-level (deoxyribonucleic acid) through whole-exome sequencing (WES) or whole-genome sequencing (WGS). Hence, it is also called genetic diagnosis. Sequencing the patient's DNA is often the first step in the diagnostic process. As each individual carries 4-5 million single nucleotide variants (SNVs) or small insertions or deletions of bases (INDELs) in its entire genome, a cascade of filtering steps need to be applied to narrow down the variant call set to the potentially disease-causing

[2] United States Congress. (2002). Rare Diseases Act of 2002. `https://www.gpo.gov/fdsys/pkg/PLAW-107publ280/html/PLAW-107publ280.htm`

[3] Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at `https://www.orpha.net`

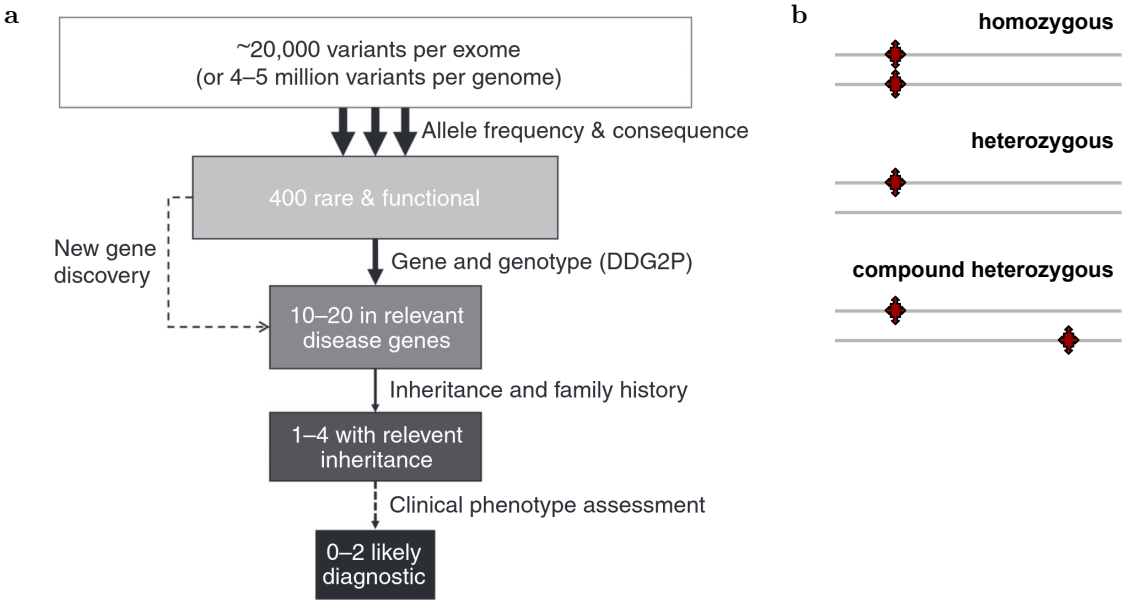

Figure 1.1: Basic variant filtering steps in NGS-based rare disease diagnostics. (a) A standard variant filtering cascade used in clinical diagnostics is presented. The aim of the filtering cascade is to narrow down the variant call set to potentially disease-causing ones by using information like allele frequency, functional consequence, clinical gene-phenotype relation, mode of inheritance. **(b)** Scheme of variant and genotypes in diploid organisms. From top to bottom: homozygous, heterozygous, and compound heterozygous. The gray line depicts the alleles of a given gene and the red dot depicts the change. Adapted from Wright et al. [2018b].

variant.[The 1000 Genomes Project Consortium, 2015; Wright et al., 2018b] Using WES instead of WGS yields up to 100,000 variants, depending on the target region, since only the coding part is sequenced, which is less than 2% of the genome. To be considered as a potential disease-causing variant, several criteria must be met, which can be directly translated into the following filter steps (Fig. 1.1a). First, variants have to be rare and should have a predicted high impact on the gene function. Second, the gene in which the variant is found has to be associated with the phenotype of interest or at least its function. Third, depending on the mode of inheritance, both alleles have to be affected (recessive) or only one allele (dominant). In the recessive case, it does not need to be the same variant (homozygous). The two variants can be at different positions as long as they affect the same gene and both alleles (compound heterozygous, Fig. 1.1b). Finally, the variants can be filtered for their segregation status, if parental information are available. Application of these filter steps typically results in 1-4 disease-causing candidate variants that are manually investigated and assessed for their clinical relevance before a molecular diagnosis is made.

It might look easy to apply 4 or 5 filters, but identifying a variant as disease-causing can be a very complex process as the underlying pathomechanism is not always clear

and can be itself complex. Interpreting genetic variants is often challenging, especially because the evidence for disease causality is not always given. Therefore, standards and guidelines for interpreting genetic variants in the clinical context of rare diseases are provided by the American College of Medical Genetics and Genomics (ACMG).[Richards et al., 2015] ACMG recommends to categorize variants based on evidence into *pathogenic*, *likely pathogenic*, *uncertain significance*, *likely benign*, and *benign* (Fig. A.1). Specifically, a nonsense or splice site variant is not enough to be classified as *pathogenic* as multiple sources of evidence have to be present. In the case of nonsense variants, it is known that they can escape nonsense mediated decay (NMD), if located in the last exon.[Popp and Maquat, 2013] Even though splice site variants lead most likely to exon skipping or altered donor/acceptor usage, it does not mean that the resulting protein product is non-functional as only a few bases in-frame could be deleted in an unimportant region. Hence, additional functional evidences are required like RNA-seq or protein quantifications.[Richards et al., 2015]

Sequence-based computational methods like VEP,[McLaren et al., 2016] SIFT,[Kumar et al., 2009] PolyPhen-2,[Adzhubei et al., 2013] CADD,[Kircher et al., 2014] MaxEntScan,[Eng et al., 2004] SpliceAI,[Jaganathan et al., 2019] and MMSplice[Cheng et al., 2019] are helpful to prioritize candidate genes in a research setting, but are not enough to establish disease causality. Here, additional curated online databases are crucial. The Genome Aggregation Database (gnomAD) stores allele frequencies of over 140,000 unrelated individuals and can be used to filter for rare or unseen variants.[Karczewski et al., 2020] ClinVar is a public database of human genetic variants and interpretations of their significance to disease that can be used to filter out benign variants or to prioritize already known pathogenic variants.[Landrum et al., 2018] Currently, it contains 1.4 million submissions from 1,880 submitters around the world for 929,054 variants in 33,122 genes.[4] Not only variant-level information can be used to prioritize variants. Also gene-phenotype associations can inform about the clinical relevance of a given variant. Online Mendelian Inheritance in Man (OMIM) is a curated public database that stores genes and phenotypes and their relationships.[McKusick, 2007] It contains 4,422 genes that cause 6,863 phenotypes, which of 5,797 are monogenic disorders.[5] The registered genes, phenotypes, and relationships increased over the years and is expected to grow further (Fig. 1.2).[Amberger et al., 2019] While Orphanet maintains a similar database with also similar numbers to OMIM, it includes relationships to orphan drugs.[3]

Overall, these resources are helpful and increase the diagnostic rate. However, the application of ACMG guidelines often leads to variant of unknown significance (VUS),[6] especially for genes that have yet to be discovered as disease causing.[Richards et al., 2015] Thus, leaving the majority without diagnosis after WES or WGS.[Neveling et al., 2013; Wortmann et al., 2015; Wright et al., 2018b; Retterer et al., 2016] Hence, comple-

[4] Accessed: 25 April 2021, `https://www.ncbi.nlm.nih.gov/clinvar/submitters/`

[5] Accessed: 25 April 2021, `https://omim.org/statistics/geneMap`

[6] A variant of unknown significance is defined as a variation in a genetic sequence for which the association with disease risk is unclear. `https://www.cancer.gov/publications/dictionaries/genetics-dictionary?cdrid=556493`

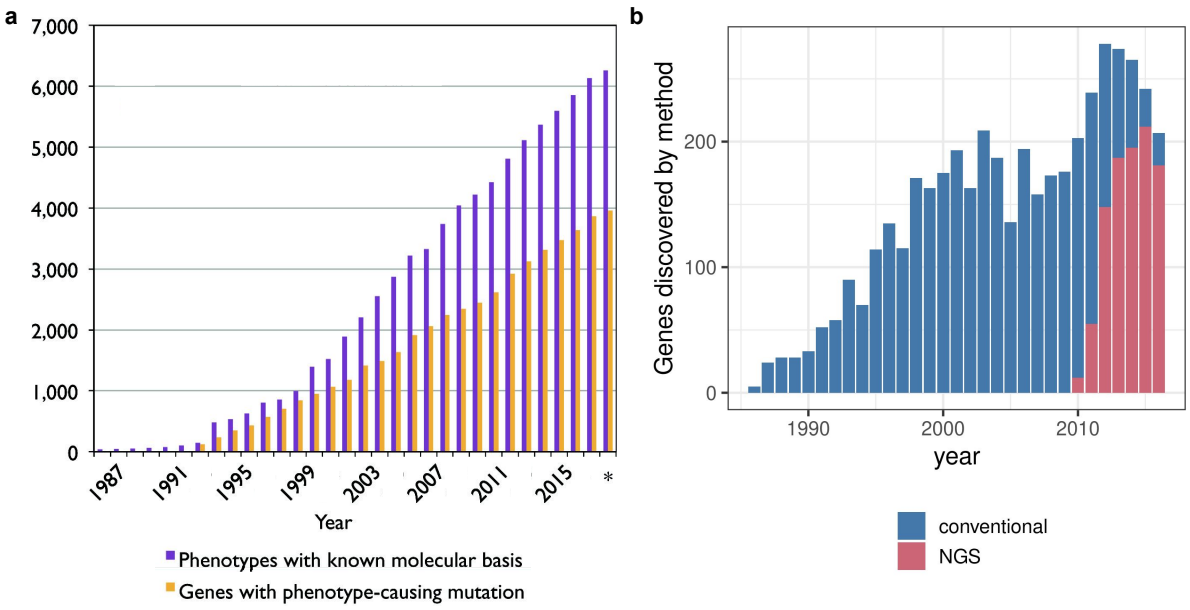

Figure 1.2: Growth of gene-phenotype relationships. The pace of disease gene discovery as cataloged by OMIM. As of 29 September 2018, there were over 6259 disorders spread across 3961 genes. **(a)** Cumulative number of registered gene and phenotypes in OMIM. **(b)** Approximate number of gene discoveries made by NGS-based approaches, WES and WGS, versus conventional approaches since 2010 Adapted from Chong et al. [2015] and Amberger et al. [2019].

mentary means to WES and WGS are needed to assess the functional impact of variants to provide additional evidence of disease causality.

1.3 DNA sequencing in diagnostics

In 1953, James Watson and Francis Crick started a new era in the filed of genomics by discovering the double-helix structure of the DNA, the blue print of our organism.[Watson and Crick, 1953] The discovery was based on the crystallized X-ray structures produced by Rosalind Franklin and Maurice Wilkins.[Zallen, 2003] Since then, researchers tried to find ways of *reading* out the DNA, our genetic information, which is build using the four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). It took till 1977 and the development of the chain-termination method by Sanger and collegues, to enable the *readout*, the so called sequencing, of the genetic code for a given locus, e.g. gene, in base-resolution.[Sanger et al., 1977] The Sanger sequencing is referred to as first generation DNA sequencing technology. In the rare disease community, Sanger sequencing was quickly adapted and used to establish molecular diagnosis but also to discover disease-causing genes and their variants (Sanger sequencing ≡ conventional sequencing, Fig. 1.2).

Concurrently with the development of the human genome project[Lander et al., 2001], the next-generation sequencing (NGS) technologies emerged.[Ronaghi et al., 1996; Mardis, 2011] With this technology the DNA is fragmented into small molecules and then sequenced massivlely in parallel. Today, with this technique, 48 complete human genomes can be sequenced in under two days.[7] As the high-throughput and cost-efficient alternative to Sanger sequencing, the NGS technology was quickly adapted by the rare disease community.[Sikkema-Raddatz et al., 2013; Bamshad et al., 2011; Gilissen et al., 2012] In diagnostics, three different NGS approaches exist: gene panel sequencing, whole-exome sequencing (WES), and whole-genome sequencing (WGS). In order to reduce sequencing costs, a targeted sequencing approach is used where the DNA is enriched for the coding regions for predifined genes (gene panel) or for all genes (WES) before sequencing. In contrast, WGS provides the full view over the genome. Despite the success of WES and WGS, Sanger sequencing continues to play an important role in diagnostics to confirm findings or to analyze rapidly and cost-effective the segregation of the variant.[Hagemann, 2015]

In 2010, WES was used for the first time to discover a disease-causing gene.[Ng et al., 2010] Since then, WES and WGS became the technology used in disease gene discovery. Already in 2015, most of the newly reported disease-gene associations were discovered through NGS (Fig. 1.2). Overall, the NGS technology revolutionized rare disease diagnostics by increasing the diagnostic yield and by accelerating discovery of novel disease genes.[Boycott et al., 2017; Neveling et al., 2013] Using WES over the last years allowed to pinpoint the causing variant in 25-50% cases.[Neveling et al., 2013; Wortmann et al., 2015; Wright et al., 2018b; Retterer et al., 2016] This number is hard to estimate and also changes depending on the underlying disorder.[Wright et al., 2018a] Despite this immense success, it also means that 50-75% of the cases are still missing a genetic diagnosis after initial WES.

1.4 Limitations of DNA sequencing in diagnostics

The continuous discovery of new gene-disease associations depicts the success of WES in rare disease diagnostics.[Chong et al., 2015; Amberger et al., 2019] Nevertheless, WES but also DNA sequencing as a whole has its shortcomings. As WES targets only the coding regions, which is less than 2% of the genome, WES is blind to the majority of variants. This limitation can be overcome by using WGS, which in principle detects most of the variants in the genome. Through its even coverage and its technical design, it even increases the sensitivity in large deletion calls. But due to our limitations in predicting the effect of and interpreting variants in non-coding regions, the 4-5 million revealed SNVs and small INDELs by WGS bring often no further insights. Hence, WGS did not increase much the diagnostic yield over WES.[Taylor et al., 2015; Clark et al., 2018; Mattick et al., 2018] Taylor et al. [2015] demonstrated in their study that 15% of the cases (5/33 variants) would have been missed by WES.

[7] Using the Illumina Novaseq 6000. `https://www.illumina.com/systems/sequencing-platforms/novaseq.html`

The reason why WES and WGS can be inconclusive is many-fold. Current challenges of these techniques include variant detection, classification, predictive power, and gene-disease as well as function-disease associations. As there is no gold standard on how to call variants, the variant detection rate can be improved through better alignments, variant calling, and filtering steps.[Boycott et al., 2017; Shamseldin et al., 2017] But even with the perfect analysis pipeline, the interpretation of the variant is the most challenging part without functional evidences. On the one hand, a variant classified as nonsense will not trigger NMD and hence not be disease causing, if it is in an isoform that is not expressed in the disease-relevant tissue.[Cummings et al., 2020] This is true for any other potentially protein changing variant class. On the other hand, a variant that is predicted to be synonymous and hence should not change the protein sequence, can still be deleterious by impacting splicing or gene expression.[Sauna and Kimchi-Sarfaty, 2011; Zeng and Bromberg, 2019] Especially for intronic and other non-coding regions, our ability to predict the impact on gene function is limited. An example are splice-effecting variants. Despite advances in sequence-based splicing prediction models through machine learning,[Xiong et al., 2015; Rosenberg et al., 2015; Cheng et al., 2019; Jaganathan et al., 2019; Cheng et al., 2021] accurate classifications remain limited. This is especially true for deep intronic variants.[Jaganathan et al., 2019] Due to this limitations in interpreting variants, the ACMG and other genetic diagnosis guidelines require additional functional evidence before a variant can be classified as pathogenic.[Richards et al., 2015; MacArthur et al., 2014] Another limitation of sequence-based predictive models is that they are based on assumptions and design decisions and are often optimized for a specific task. In the case of splicing models, this leads to a high rate of missing predictions or ignored variants, especially for deep intronic variants.[Jian et al., 2014] Another problem is the candidate gene prioritization based on phenotypic data. As genes are filtered based on disease association or functional relevance, extensive annotations are required but also a comprehensive phenotypic characterization of the patient. Both are often incomplete or even missing as the continuum of pathologies in rare diseases are difficult to objectively segment into discrete disease entities.[Boycott et al., 2017]

Overall, these limitations highlight the need for alternatives but foremost complementary technologies in rare disease diagnostics. The many omics research fields offer a variety of promising technologies that need to be systematically investigated for their potential to further improve the diagnostic yield. These include especially transcriptomics, proteomics, and metabolomics, among others.

1.5 RNA sequencing in diagnostics

Like DNA, ribonucleic acid (RNA) is one of the four major macromolecules essential to all known life forms. RNA is a chain of nucleotides similar to DNA. The main differences to DNA is the use of ribose in the backbone, the use of uracil (U) instead of thymine, and that it is found mostly single stranded in nature. In cellular organisms, genes are translated into messengers RNA and then used as template to synthesize proteins

by translating the codons into amino acids. The building block for RNA-seq was laid in 1971, when the reverse transcriptase was discovered.[Gallo, 1971] Using the reverse transcriptase, Shaffer et al. [1990] were the first to measure gene expression by using the reverse transcription-polymerase chain reaction (RT-PCR) method. Shortly after NGS was established for DNA sequencing, the protocols were adapted for sequencing the RNA in the early 2000s.[Weber, 2015] The main steps for RNA-seq are converting RNA to DNA using the reverse transcriptase followed by a amplification of the DNA molecules through polymerase chain reaction (PCR) followed by NGS. Depending on the application an enrichment step in the beginning can be used to discard ribosomal RNA or target specific transcripts (Fig. 1.3).[Wang et al., 2009] It is important to keep these steps in mind, as they are sources of technical biases like lower abundance of transcripts having a high G/C content or containing long homopolymer stretches.[Risso et al., 2011; Hansen et al., 2012; Weber, 2015] Despite the technical biases, RNA-seq superseded technologies like RT-PCR or microarrays[Schena et al., 1995] as RNA-seq has multiple advantages: (i) it does not depend on prior sequence knowledge, (ii) it provides a direct measure of RNA abundance and splicing, and (iii) simultaneous sequence discovery including variant calling and quantification.[Weber, 2015]

One of the first studies using RNA-seq were done on prostate cancer cell lines and plants starting in 2006.[Bainbridge et al., 2006; Cheung et al., 2006; Emrich et al., 2007; Weber et al., 2007] In diagnostics, RNA-seq was initially used for validation or as further evidence rather than as a primary detection tool.[Wang et al., 2013; Van Keuren-Jensen et al., 2014] Later it was also used in single gene studies to detect aberrant events.[Chandrasekharappa et al., 2013; Kernohan et al., 2017] But prior to this work, no systematic study on the utility of RNA-seq as a diagnostic tool in rare diseases has been performed. Since 2017, when two studies independently and in parallel demonstrated the utility and complementarity of RNA-seq to WES, RNA-seq has been increasingly used in the diagnosis of rare inherited diseases (Fig. 1.4).[Cummings et al., 2017; Kremer et al., 2017] One of the two pioneering studies is part of this thesis. Over the course of time, RNA-seq proved to increase the diagnostic rate in WES or WGS inconclusive cases by 10-36%.[Cummings et al., 2017; Kremer et al., 2017; Gonorazky et al., 2019; Frésard et al., 2019; Lee et al., 2019; Maddirevula et al., 2020; Rentas et al., 2020; Murdock et al., 2021] The bioinformatic methods and approaches used throughout the studies differed considerably, but ultimately RNA-seq was used to find three different classes of events: aberrant gene expression, aberrant splicing, and monoallelic expression of the rare allele. The latter is also referred to as allelic imbalance.[Cummings et al., 2017; Mohammadi et al., 2019] The proportion of classes detected in the studies differed, which can be attributed to the underlying disease but also to the use of the methods. In addition, Gonorazky et al. [2019] successfully used RNA-seq to call variants by identifying the disease-causing variant previously missed by WES. The relevant methods and studies (Fig. 1.4) are introduced and discussed within this work in the appropriate sections (Section 2, 3, and 4, 6).

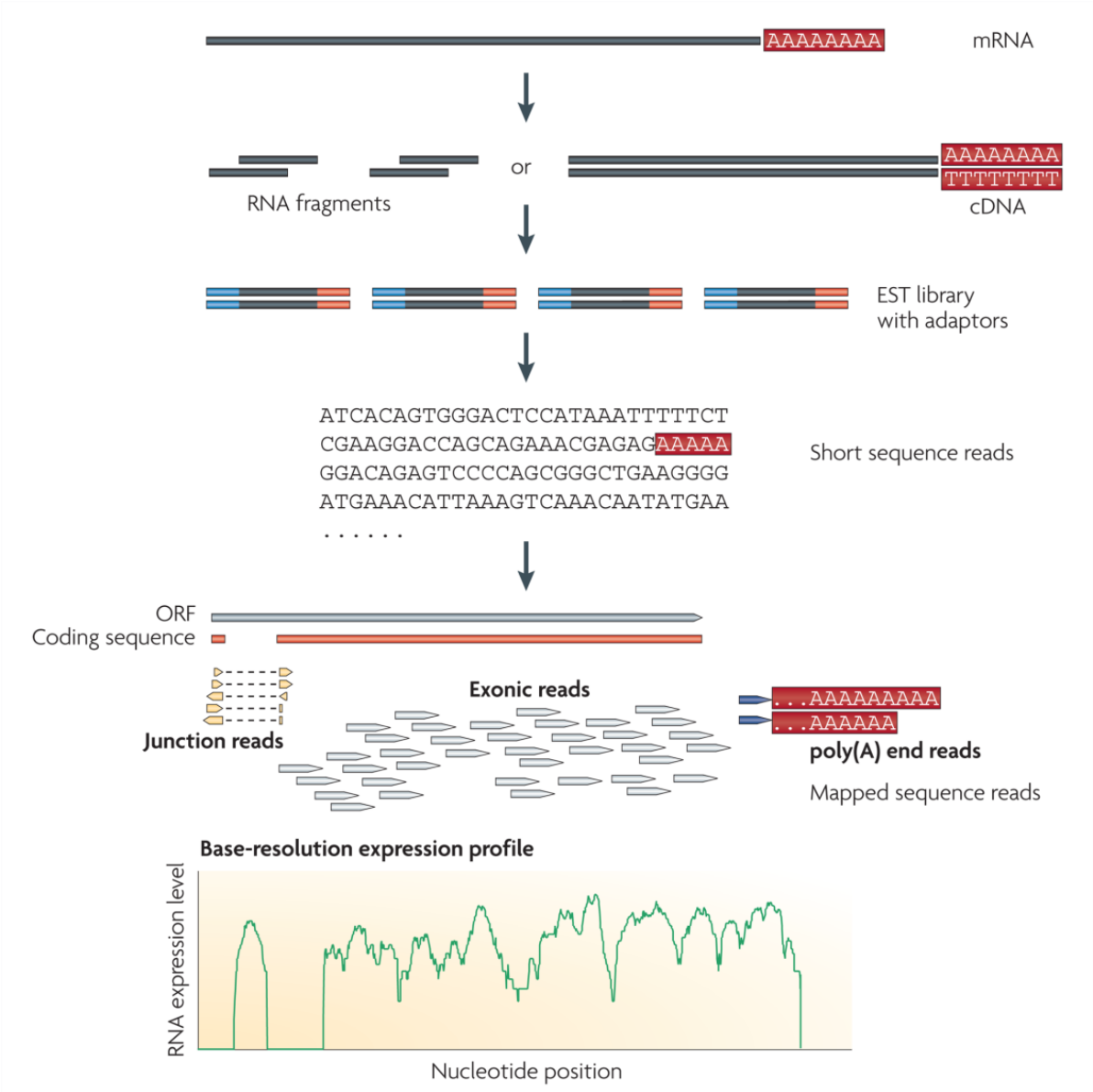

Figure 1.3: A typical RNA-Seq experiment Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (see main text). Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown. Taken from Wang et al. [2009].

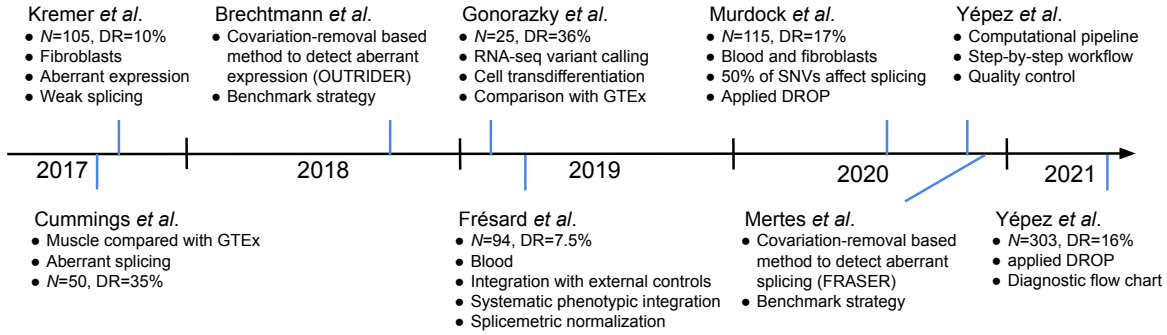

Figure 1.4: Timeline of studies that advancing RNA-seq-based rare disease diagnostics. The timeline depicts the relevant studies in the field of RNA-seq-based rare disease diagnostics together with their main contribution and methodology. Out of the 9 studies 3 focused on the development of specialized methodologies to detect aberrant events in RNA-seq data. The other 6 studies focused on using RNA-seq to diagnose WES and WGS inconclusive cases. N: number of samples, DR: diagnostic rate. Courtesy of Vicente Yépez.

1.6 Quantification of gene and splicing metrics

Gene expression quantification The simplest thing to do with RNA-seq is to quantify gene expression. This essentially amounts to counting reads that are aligned within a region in the genome. The region can be the gene body or only the exonic part of a gene. As genes can overlap and reads can span multiple genomic features (e.g. exons) or extent into the intron or even map to multiple regions, decisions have to be made on how and what should be counted. Over the years multiple tools where developed to quantify gene expression in RNA-seq data.[Li and Dewey, 2011; Anders et al., 2015; Bray et al., 2016; Patro et al., 2017] Quantifying gene expression usually results in a count matrix of size $p \times N$ where the rows are the genes or genomic features (p) and the columns are the samples (N). Kallisto[Bray et al., 2016] and Salmon[Patro et al., 2017] use the pseudoalignment and quasi-mapping idea where no alignment is needed and gene expression is estimated by mapping parts of each read against a pre-build index based on the transcriptome. To reduce noise and signal from premature mRNAs, reads that overlap intronic parts or align to multiple loci are usually discarded by the traditional methods.

Splicing quantification Compared to gene expression quantification where reads are counted falling within the boundary of the gene or exon, splicing can be quantified in two ways: (i) globally at the isoform level per gene or (ii) locally at the exon or intron level. Regardless of the quantification methods, reads that span exon-exon boundaries, so called split reads, are used mainly as evidence of splicing.[Sultan et al., 2008] Quantifying isoforms is a challenging task because the short reads generated by RNA-seq usually can not be uniquely assigned to an isoform. Therefore, heuristics or statistical models must be applied to estimate isoform expression.[Li and Dewey, 2011; Trapnell et al., 2010;

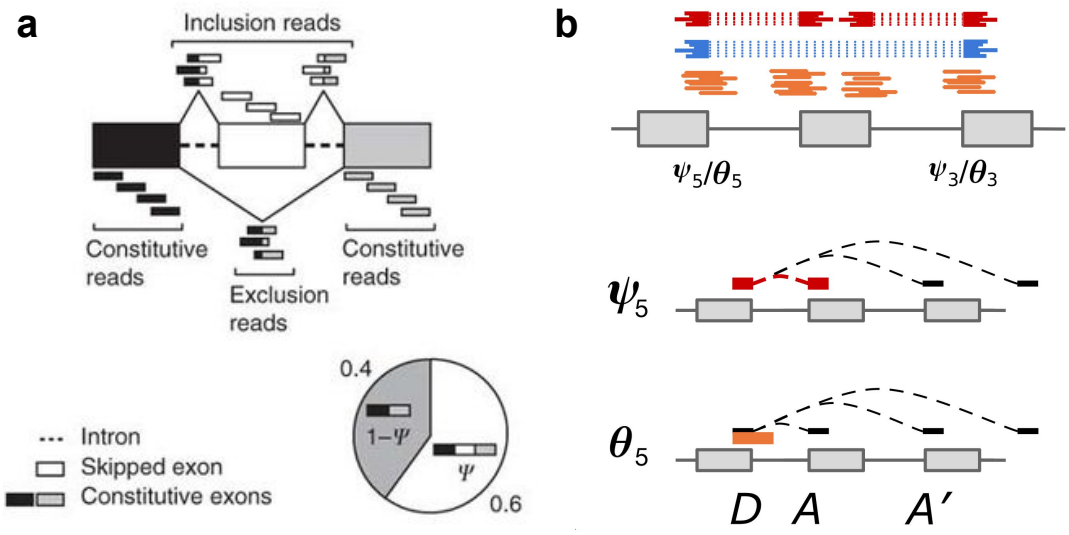

Figure 1.5: The exon- and intron-centric percent spliced in (ψ) metric (a) The exon-centric percent spliced in metric Ψ is defined as the number of reads supporting exon inclusion as the fraction of the combined number of reads supporting inclusion and exclusion.[Katz et al., 2010] It requires the gene model a priori to know which split reads to consider. **(b)** The intron-centric 5' percent spliced in value (ψ_5) is calculated purely based on split reads as the number of reads supporting the splicing event from D to A relative to the combined number of reads supporting splicing from D to any acceptor site A'. The splice-site-centric donor splicing efficiency (θ_5) uses the the non-spliced reads overlapping the donor site over the full coverage at the donor, total number of split and non-spliced reads. The ψ_3 and θ_3 is calculated analogously. The intron or splice site of interested is colored in red and orange, respectively. Adapted from Katz et al. [2010] and Pervouchine et al. [2013].

Roberts and Pachter, 2013; Patro et al., 2014, 2017] Alternative splicing can also be quantified on the local level in an exon-centric view by calculating the inclusion of an exon (Fig. 1.5a).[Katz et al., 2010; Tilgner et al., 2012] Katz et al. [2010] defined the percent spliced in Ψ value as the number of reads supporting exon inclusion as the fraction of the combined number of reads supporting inclusion and exclusion. This metric focuses on a particular splicing pattern, namely the skipping of a single exon. This requires a priori knowledge of the gene model to identify the exon-exon boundaries used for quantification. It also simplifies the biology of splicing by ignoring potentially additional exon-exon boundaries.

An annotation-free quantification of splicing can be achieved with the intron-centric splicing metrics 5' and 3' splicing index (ψ_5 and ψ_3), also called 5' and 3' percent spliced in, which is related to the percent spliced in value (Fig. 1.5b). Specifically, Pervouchine

et al. [2013] define the intron-centric splicing metrics as:

$$\psi_5(D,A) = \frac{\mathrm{n}(D,A)}{\sum_{A'} \mathrm{n}(D,A')} \quad \text{and} \tag{1.1}$$

$$\psi_3(D,A) = \frac{\mathrm{n}(D,A)}{\sum_{D'} \mathrm{n}(D',A)}, \tag{1.2}$$

where D is a donor site and A is an acceptor site. $\mathrm{n}(D,A)$ denotes the number of reads spanning the given intron from D to A. The summands in the denominators are computed over all acceptors A' that spliced with the donor D of interest (Eq. 1.1) and all donors D' that spliced with the acceptor A of interest (Eq. 1.2) Assuming uniform read coverage across the gene, one can estimate the exon-centric version as:[Pervouchine et al., 2013]

$$\Psi = \frac{\psi_5 + \psi_3}{2}. \tag{1.3}$$

The percent spliced in values only consider alternative splicing events and ignore by design the splicing efficiency. Splicing efficiency is calculated by computing the ratio of reads that were not spliced and hence align to both sides of a given splice site over split reads similar to ψ_5 and ψ_3 (Fig. 1.5). Multiple related definitions exist including 3' splice site ratio,[Khodor et al., 2011] completeness of splicing index,[Tilgner et al., 2012] and percent intron retained.[Braunschweig et al., 2014] Pervouchine et al. [2013] defined splice-site-centric splicing efficiency as:

$$\theta_5 = \frac{\sum_{A'} \mathrm{n}(D,A')}{\mathrm{n}(D) + \sum_{A'} \mathrm{n}(D,A')} \quad \text{and} \tag{1.4}$$

$$\theta_3 = \frac{\sum_{D'} \mathrm{n}(D',A)}{\mathrm{n}(A) + \sum_{D'} \mathrm{n}(D',A)}, \tag{1.5}$$

where $\mathrm{n}(D)$ is the number of non-split reads spanning the exon-intron boundary of donor D, and $\mathrm{n}(A)$ is defined as the number of non-split reads spanning the intron-exon boundary of acceptor A.

1.7 Outlier detection

Outlier detection is the process of finding observations in the data that are significantly deviating from other observations. This process is also known as novelty detection or anomaly detection. Over the last centuries many different methods and criteria for outlier classification have been proposed.[Chauvenet, 1863; Dean and Dixon, 1951; Grubbs, 1969; Cook, 1977; Hodge and Austin, 2004] As early as 1863, William Chauvenet proposed the Chauvenet criterion that is based on defining a probability band in which data points should be lying based on the normal distribution.[Chauvenet, 1863] This is similar to the widely used z score approaches. Today outlier detection is an active

research field and has practical use cases in many real world problems like credit card fraud, machine failure in production pipelines, statistical research, and sport event analysis.[Zimek and Filzmoser, 2018] It is used mainly in two ways: (i) the detection of an aberrant event in the data in order to remove it, learn from it, or act on it and (ii) to robustly estimate distribution parameters by removing or downweighting the influence of the outlier event, if present, on the model. These two classes are also called accommodation methods and discordancy tests.[Barnett and Lewis, 1974] Since outlier detection is based on an a priori specified distribution or model, one cannot be sure whether the assumptions, and thus the underlying model, need to be changed or the observation is a *true* outlier. Therefore, classifying an observation as an outlier is ultimately a subjective task.[Collett and Lewis, 1976; Zimek and Filzmoser, 2018; Leys et al., 2019]

A simple and widely used approach to identify outliers is the usage of a z score cutoff. This assumes that the underlying data is based on a normal distribution. z scores are defined based on the mean $\bar{x}$ and standard deviation $\sigma(x)$ of all observations and is defined as:

$$z_i = \frac{x_i - \bar{x}}{\sigma(x)}. \tag{1.6}$$

Outlier data points are then identified by choosing a cutoff that is usually 2 or 3 standard deviations away from the mean ($|z_i| > 2$ and $|z_i| > 3$, respectively). However, this is a subjective and arbitrary definition with no assessment of the significance of the respective event. Alternative ways that assesses the significance of an outlier exists like the Grubbs test[Grubbs, 1969] or Dixon's Q test.[Dean and Dixon, 1951] Further, in differential gene expression analysis, to obtain robust estimates, suspect observations are completely excluded during model fitting based on Cook's distance[Love et al., 2014] or their influence on the model is reduced with weights based on Pearson residuals.[Zhou et al., 2014]

When measuring gene expression data across samples, outlier data points can be defined on the gene level across the samples or as a whole for a given sample across all genes. This refers to the difference between univariate and multivariate outlier detection, respectively. While the multivariate case is interesting for testing whether an experiment has failed for a given sample, the univariate case is interesting for the use in rare diseases where only a single event is expected in a given sample, while the majority of the remaining genes are assumed to have similar expression levels compared to the population. Even though, outliers are detected in both cases, the underlying methods differ. While outliers in the univariate case are detected as values that differ significantly from a robust central tendency estimator (the mean in the case of z scores), an ellipse in the 2-dimensional space or a complex multidimensional cloud in a high dimensional space has to be modeled before the outlier can be detected.[Cousineau and Chartier, 2010; Leys et al., 2019] To calculate the distance of a given data point in a multidimensional space to the centroid, a cloud defined by the majority of observations in the data, the Mahalanobis distance is often used.[Mahalanobis, 1930; Leys et al., 2019; Filzmoser and Gregorich, 2020] Similar to the z score, the Mahalanobis distance

does not provide an assessment of the significance of the finding. To assess this, empirical P values need to be calculated, as recently done by Ferraro et al. [2020] to detect splice outliers. Alternatively, if the underlying data of each variable follows a Student's t distribution, Hotelling's T-squared distribution (T^2) can be used to detect outliers in a multidimensional space.[Hotelling, 1931]

1.8 Autoencoders

An autoencoder is an artificial neural network that is used to learn efficient encodings of high dimensional data. By learning a small and compressed representation of the data, also called the latent space or encoding, it extracts essential features. Thus, it can be used for dimensionality reduction. Autoencoders were introduced around 1990 by Lecun [1987]; Bourlard and Kamp [1988]; Hinton and Zemel [1994]. The autoencoder has to sides, the encoder which maps the input data to the representation and a decoder which reconstructs the data using the compressed representation (Fig. 1.6a). Specifically an autoencoder is defined by the two functions:

$$f_\theta : \mathbf{X} \rightarrow \mathbf{H}, \text{and} \tag{1.7}$$

$$g_{\theta'} : \mathbf{H} \rightarrow \hat{\mathbf{X}}, \tag{1.8}$$

where $\mathbf{X}$ is the input matrix, $\mathbf{H}$ the latent space, and $\hat{\mathbf{X}}$ the reconstruction. The parameters θ and θ' of the encoder and the decoder, respectively, are optimized by minimizing the reconstruction error. To force the autoencoder to learn a true representation instead of the identity matrix, the latent space is usually magnitudes smaller than the input, which represents the so called bottle neck. The encoder and decoder can each be composed of multiple layers with non-linear function and hence is in principle capable of learning representations for high dimensional and complex data. Thus, autoencoders and its variations can be applied to various data types from language processing,[Grozdic and Jovicic, 2017] image processing,[Dai and Wang, 2018] object detection,[Park et al., 2018a] and biometric recognition.[Yu et al., 2018] As example, Way and Greene demonstrated that autoencoders can be used to extract meaningful biological features from gene expression data.[Way and Greene, 2018]

Denoising autoencoders Initially, autoencoders were used to learn compressed representations in an unsupervised fashion. But recent research showed that they can be used to reconstruct or denoise input data. By adding noise to the input data before mapping it to the latent space, autoencoders are trained to recover the original data by removing the noise (Fig. 1.6b).[Vincent et al., 2008] This subclass of autoencoders are so called denoising autoencoders. Noise functions can be arbitrary, but three types of noise are usually applied depending on the use case: (i) adding Gaussian noise, (ii) masking

[8] Arden Dertat, Oct 3, 2017; Accessed on 29. April 2021: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

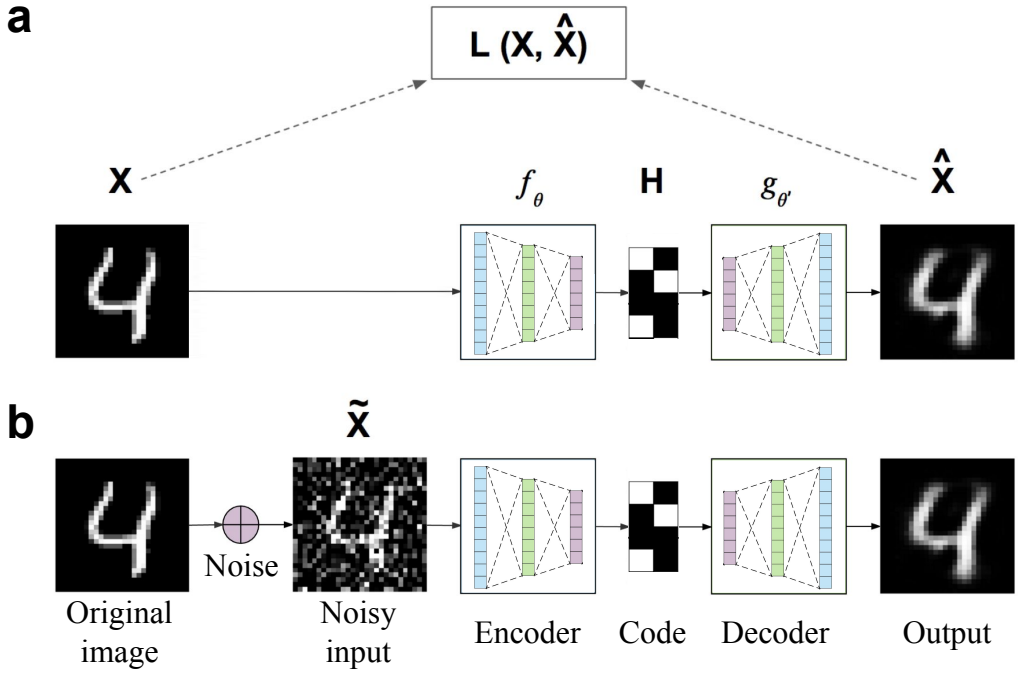

Figure 1.6: Schematic architecture of a (denoising) autoencoder. **(a)** Usage of a simple autoencoder to learn the encoding of numbers. The input ($\mathbf{X}$) is mapped with the encoder (f_θ) to the latent space ($\mathbf{H}$) and reconstructed to $\hat{\mathbf{X}}$ by the decoder (g_θ). **(b)** Adding noise to the input before mapping it to the latent space turns the autoencoder into a denoising autoencoder. In both approaches, the parameters of the encoder and decoder are optimized by minimizing the error between the original input and the reconstruction ($\mathrm{L}(\mathbf{X}, \hat{\mathbf{X}})$). Adapted from Arden Dertat[8] with input from Stefan Loipfinger.

a fraction of data points by setting them to zero, and (iii) setting a fraction of data points to their minimum or maximum at random, which is known as salt-and-pepper noise.[Vincent et al., 2010] In single cell genetics, Eraslan et al. [2019] and Badsha et al. [2020] showed successfully how such denoising autoencoders can be applied to denoise gene expression data and to impute missing values.

1.9 Aims and scope of this thesis

The overall aim of this thesis is to improve the genetic diagnosis of rare diseases through the detection of aberrant events in RNA-seq data. The contribution of my work is four-fold: (i) a systematic evaluation of the utility of RNA-seq in rare disease diagnostics, (ii) the development of a robust gene expression outlier detection method, (iii) the development of a robust aberrant splicing detection method, and (iv) the finding of new disease-causing events in 3 cases leading to a molecular diagnosis.

Systematic evaluation of the utility of RNA-seq in rare disease diagnostics. At the beginning of this work, RNA-seq was used mainly in clinical studies as validation or in single-gene studies. But a systematic evaluation of its utility for rare disease diagnostics in general and for WES or WGS inconclusive cases in particular, was missing.

Therefore, together with my colleagues, I investigated the usability of RNA-seq in rare disease diagnostic, by developing tools to detect aberrant gene expression, aberrant splicing events and mono-allelic expression (MAE) of the alternative allele. By providing 5 new genetic diagnoses in WES inconclusive cases, I proved its power and complementarity to WES and WGS. Further, I showed that weak splicing is a frequent cause of aberrant expression of cryptic exons caused by rare variants effecting splicing.

Detection of aberrant gene expression with OUTRIDER As RNA-seq-based rare disease diagnostic is a relative new research field, no specialized method existed to detect gene expression outliers. As the first part of this work highlighted together with other studies the need for robust detection methods, I, together with Felix Brechtmann and Agnė Matusevičiūtė, developed OUTRIDER (OUTlier in RNA-seq fInDER), an autoencoder-based gene expression outlier detection method to fill this gap. To evaluate OUTRIDER, I developed new benchmark strategies using the Genotype-Tissue Expression (GTEx) dataset.[The GTEx Consortium et al., 2015]

Detection of aberrant splicing events with FRASER The initial studies, which used RNA-seq to detect aberrant splicing, highlighted the needs for improved outlier detection. In particular, it became evident that splicing data must be controlled for confounding effects and that intron retention has to be taken into account.

To this end, together with Ines Scheller, I developed FRASER (Find RAre Splicing Events in RNA-seq), which follows the same principles of OUTRIDER. It uses an autoencoder to control the data for confounders and uses a beta-binomial (BB) distribution to identify aberrant splicing events. Again I developed benchmarking strategies utilizing the GTEx dataset.[The GTEx Consortium et al., 2015]

Solving WES inconclusive cases using RNA-seq While being developed, OUTRIDER and FRASER were continuously applied on new incoming data. This provided in total over 40 new diagnoses over the last years. The result of this ongoing study was published recently by Vicente Yépez, Mirjana Gusic, and colleagues.[Yépez et al., 2021a] In this part, I will highlight some interesting cases, to showcase the importance and complementarity of RNA-seq in rare disease diagnostics.

2 Exploring the utility of RNA sequencing in rare disease diagnosis

> *With rare disease research, we are not following a well-trodden path; we're making the path. There is usually no effective standard of care and no drug has gone through the regulatory process.*
>
> Phil Vickers, Ph.D., 2015

The methodology, results, and figures presented in this chapter are part of the manuscript "Genetic diagnosis of Mendelian disorders via RNA sequencing" from Kremer et al. [2017]. The author's contributions are included in it. In short, I performed the splicing analysis, weak splice site modelling and mono-allelic expression analysis with the help of Daniel Bader and the supervision of Julien Gagneur.

2.1 Motivation

Over the last decade, the NGS technology transformed the way diagnostics is done in the field of Mendelian disorders by increasing the diagnostic yield and by accelerating discovery of novel disease genes.[Boycott et al., 2017; Neveling et al., 2013] Sequencing the full coding part of the genome in a high throughput manner allowed to pinpoint the causing variant in 25-50% cases.[Neveling et al., 2013; Wortmann et al., 2015; Wright et al., 2018b; Retterer et al., 2016] Despite this immense success, it also means that 50-75% of the cases are still missing a genetic diagnosis after initial WES. Even though WES might detect the disease-causing variant, the variant is often not prioritized due to the labeling as a VUS[1]. VUS classification is not the only reason a WES can be inconclusive. Current challenges include variant detection, predictive power of sequence-based algorithms, prioritization, and interpretation. As WES only covers 2% of the genome WGS can in principle overcome some shortfalls of WES by detecting most of the variants in the genome. But due to the sheer amount of variants WGS is revealing and our limitations in predicting the effect of and interpreting variants in non-coding

[1] A variant of unknown significance is defined as a variation in a genetic sequence for which the association with disease risk is unclear. `https://www.cancer.gov/publications/dictionaries/genetics-dictionary?cdrid=556493`

regions, WGS did not increase much the diagnostic yield over WES.[Taylor et al., 2015; Clark et al., 2018; Mattick et al., 2018]

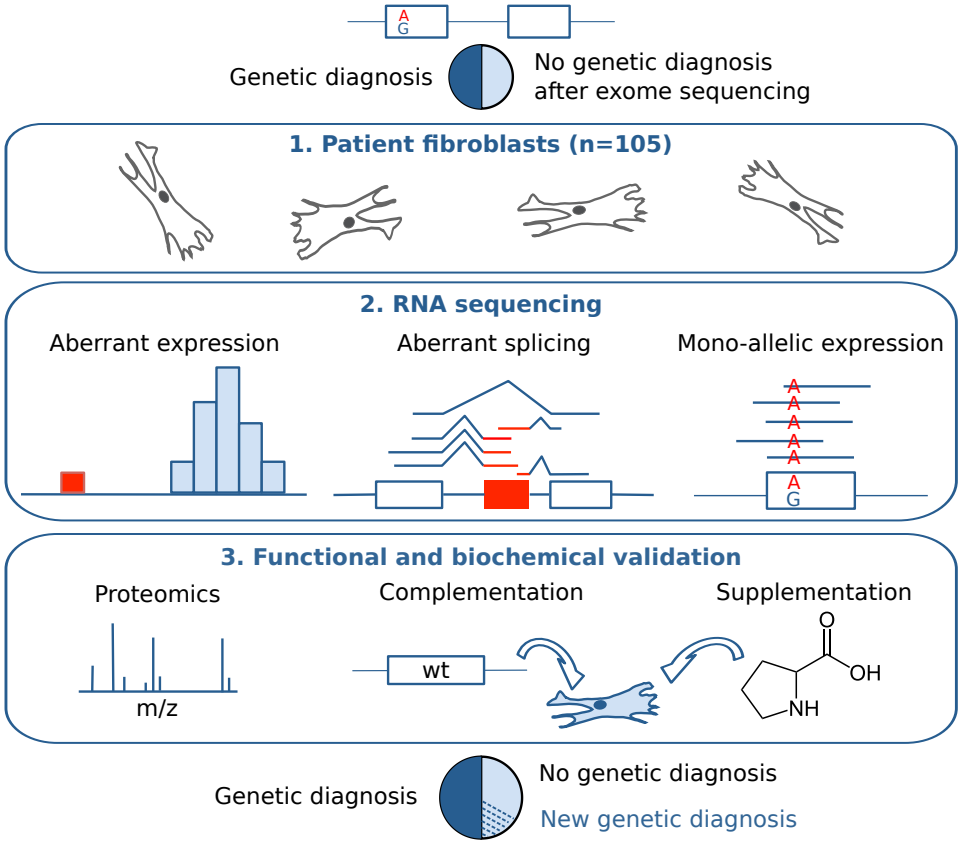

Figure 2.1: Strategy for genetic diagnosis using RNA-seq. The approach we followed started with RNA-seq of fibroblasts from unsolved WES patients. Three strategies to facilitate diagnosis were pursued: Detection of aberrant expression (for example, depletion), aberrant splicing (for example, exon creation) and mono-allelic expression of the alternative allele (for example, A as alternative allele). Candidates were validated by proteomic measurements, lentiviral transduction of the wild-type (wt) allele or, in particular cases, by specific metabolic supplementation. Taken from Kremer et al. [2017].

With this study, we aimed to overcome some of the limitations of the solely sequence-based approaches by probing functional consequences of genetic variants using RNA-seq. As RNA-seq captures gene expression, splicing patterns, as well as allele specific expression, it is complementary to the WES and WGS approaches and can directly help to interpret the effect of a given variant and can even pinpoint the underlying pathomechanism in some cases. Here, we focused on three extreme situations to prioritize putative disease-causing genes for rare diseases (Fig. 2.1): (i) aberrant gene expression levels, where the gene counts lie outside the normally observed range, (ii) aberrant splicing events, where e.g. a cryptic exon is aberrantly expressed due to a deep-intronic

SNV, (iii) MAE of the allele harbouring the rare variant, where the other allele can be silenced due to e.g. a heterozygous deleterious variant in the promotor region leaving only the observed allele expressed.

2.2 Evaluating RNA-seq based diagnostics in a mitochondrial disease cohort

By the time I started my PhD in 2016, RNA-seq was not yet evaluated in a systematic study nor its power in facilitating molecular diagnosis was accessed. Even sophisticated tools to detect aberrant events in RNA-seq data did not exist. Hence, we considered investigating the power of RNA-seq for molecular diagnosis with a panel of patients diagnosed with a mitochondrial disorder and to prototype tools to guide the diagnosis. We started our study with a mitochondrial disease cohort as it has three advantages for such a pilot study: (i) collectively it represent one of the most frequent inborn errors of metabolism affecting 2 in 10,000 individuals,[Gorman et al., 2016] (ii) the broad range of unspecific clinical symptoms and the genetic diversity in mitochondrial diseases makes molecular diagnosis difficult with more than 300 disease-causing genes and WES often resulting in VUS,[Mayr et al., 2015; Wortmann et al., 2017] (iii) the access to fibroblast cell lines in the patient cohort and the ease to validate candidate variants by perturbation and complementation assays in fibroblasts.[Haack et al., 2010].

For this pilot study, we selected 105 patients with suspected mitochondrial disease in whom fibroblast cell lines were available. In 48 times of these cases, the WES was inconclusive and failed to provide a genetic diagnosis. The other 57 cases received a molecular diagnosis after WES and were included in the study to validate the performance and demonstrate the added value of RNA-seq. In short, the 119 fibroblast cell lines from the 105 cases included 6 controls and 8 replicates were subjected to non-strand specific, polyA-enriched RNA-seq. After initial RNA isolation and quality controls, the library was prepared as described in the Low Throughput protocol of the TruSeq RNA Sample Prep Guide (Illumina). The RNA libraries were then sequenced as 100 bp paired-end runs on an Illumina HiSeq2500 platform. After demultiplexing, the FASTQ files were mapped against the hg19 genome assembly[Casper et al., 2018] with STAR[Dobin et al., 2013] (version 2.4.2a). To increase the detection of novel introns and gene fusions the following two parameters were changed from the defaults: *chimSegmentMin = 20* and *twopassMode = Basic*. The downstream analysis was restricted to properly aligned read pairs (read mates from opposite strands), the standard chromosomes 1-22, X, Y, and M, as well as to the 27,682 UCSC Known Genes.[Hsu et al., 2006] A more detailed description of the method can be found in Kremer et al. [2017]. In the this thesis, this datasets will be referred to as the Kremer dataset.

The next sections describe in detail the three strategies we developed and then used to systematically prioritize potential disease-causing genes in RNA-seq data: (i) detection of aberrant gene expression levels (2.2.1); (ii) detection of aberrant splicing patterns (2.2.2); (iii) detection of mono-allelic expression of an alternative rare variant (2.2.3).

The sections are divided into two parts, the first describing the methodology and methods, followed by the second describing the results.

2.2.1 Detection of aberrant gene expression levels

The aberrant expression analysis was mainly performed by Daniel Bader and described in Kremer et al. [2017]. Nevertheless, I will include it in this thesis for completeness.

Starting with the binary alignment map (BAM) files, we considered any read pair that overlapped completely a given gene body on either strand orientation. We used the *summarizeOverlaps* function provided by the R/Bioconductor *GenomicAlignments*[Lawrence et al., 2013] package to extract the gene count k_{ij} of the gene $j = 1, \ldots, p$ in sample $i = 1, \ldots, N$. While counting the reads, we used the following parameter settings: *mode = intersectionStrict*, *singleEnd = FALSE*, *ignore.strand = TRUE*, and *fragments = FALSE*. In order to remove noisy, non-expressed, and non-detected genes, we filtered out genes, if their 95^{th} percentile of fragment counts across all samples were below 10. After quality control, we identified 12,680 transcribed genes in 119 RNA-seq samples (Fig. A.2) and got a gene count matrix $\mathbf{k}$ with the dimension $p \times N$. Hierarchical clustering of the resulting count matrix revealed three top level clusters that could not be linked to any biological or technical properties of the samples. Therefore, we considered them as technical variation of unknown origin (Fig. A.3). As the samples were taken from different body parts, we corrected additionally the counts using the 5 most viable *HOX* genes as they are important regulators in the development of the human body parts.[Lewis, 1978]

To detect aberrant gene expression levels, we adapted the methodology from DESeq2[Love et al., 2014] an R/Bioconductor package developed for the purpose of differential gene expression analysis. Instead of comparing two groups of samples as it is done in differential gene expression, we compared one individual against the rest of the cohort. Specifically, we modelled the read count k_{ij} with a generalized linear model:

$$\begin{aligned} K_{ij} &\sim \mathrm{NB}(s_i \times q_{ij}, \alpha_j) \\ \log_2(q_{ij}) &= \beta_j^0 + \beta_j^{\text{condition}} x_{ij}^{\text{condition}} + \beta_j^{\text{batch}} x_{ij}^{\text{batch}} + \beta_j^{\text{sex}} x_{ij}^{\text{sex}} + \beta_j^{\text{hox}} x_{ij}^{\text{hox}}, \end{aligned} \tag{2.1}$$

with NB being the negative binomial distribution, α_j the dispersion parameter of gene j, s_i the size factor of sample i, and β_j^0 the intercept parameter for gene j. The value of $x_{ij}^{\text{condition}}$ was set to 1 for all RNA samples i of the case of interest, thereby allowing for biological replicates, and 0 otherwise. The resulting value $\beta_j^{\text{condition}}$ represents the $\log_2$-fold change of gene j for a given case against all others. The z scores were computed by dividing the fold changes by the standard deviation of the normalized expression levels of the respective gene. Finally, the negative binomial P values were corrected for multiple testing per sample using Hochberg's family-wise error rate (FWER) method.[Hochberg, 1988]

Overall, only a few aberrant gene expression events were detected. More specifically, 1 event was detected in median, whereas 90% of samples had < 10 events and only 4

samples had > 100 events with a large effect (|z score| > 3) and significant differential expression (Hochberg adjusted P value < 0.05, Fig. 2.2). Among the most aberrantly expressed genes across th WES inconclusive cases, we found 2 genes encoding mitochondrial proteins, *MGST1* (one case) and *TIMMDC1* (two cases) to be significantly down-regulated (Fig. 2.2b–d). In each case, WES did not identify any variants in the respective gene that could explain this down-regulation. Quantitative proteomics was used to validate but also evaluate the consequences of the down-regulation. Patient #73804 showed $\sim$ 2% of control MGST1 level, whereas the lack of detection of TIMMDC1 in both patients (#35791 and #66744) confirmed an even stronger effect on protein expression, indicating loss of function (Kremer et al. [2017] Fig. 2e and supplemental Fig. 4).

MGST1, a microsomal glutathione S-transferase, is involved in the oxidative stress defense.[Lee et al., 2008] Despite the fact that WES and subsequent WGS analysis failed to detect a rare variant in *MGST1* that could explain the effect at the RNA and protein level, it provides the likely cause of the infantile-onset neurodegenerative disorder.[Holzerova et al., 2016] The evidence of down-regulation of *MGST1* even suggests treatment with antioxidants.

Both cases with the aberrant down-regulation of *TIMMDC1* presented with muscular hypotonia, developmental delay and neurological deterioration, which led to death in the first 3 years of life. Quantitative proteomics analysis showed not only the lack of TIMMDC1 but also reduction of complex I sub units in fibroblasts, which is consistent with the described function of TIMMDC1 as a respiratory chain complex I assembly factor.[Guarani et al., 2014; Andrews et al., 2013] In addition, we were able to increase the levels of complex I sub units through re-expression of TIMMDC1 wildtype. Altogether, this underlines the disease-causing effect of TIMMDC1-deficiency and the important role of TIMMDC1 in the assembly of the complex I. The identification of homozygous deep intronic rare variants in *TIMMDC1* in both cases through the analysis of the RNA-seq data let us to the aberrant splicing analysis.

2.2.2 Detection of aberrant splicing

We followed the same idea as in the aberrant gene expression analysis (Section 2.2.1). We adapted LeafCutter, which was developed to detect differential splicing,[Li et al., 2018] to find aberrant splicing events in RNA-seq data by testing each patient against the rest of the cohort. Leafcutter is designed as an annotation-free algorithm and as such can detect splice sites de novo. But due to this design of only looking at split reads, LeafCutter is intrinsically blind to intron retention events and therefore we put our focus on alternative splicing in this analysis. First, we modified the split read counting and clustering parameters in order to detect rare clusters, capture local gene fusion events and to detect sample-specific introns. Specifically, we used *minclureads = 30*, *maxintronlen = 500,000*, and *mincluratio = 1e-5*. After filtering and clustering of the split read counts, each sample was tested against the rest of the cohort using *min_samples_per_group = 1* and *min_samples_per_intron = 1*. Finally, the resulting Dirichlet-Multinomial-based P values were corrected for multiple testing per sample us-

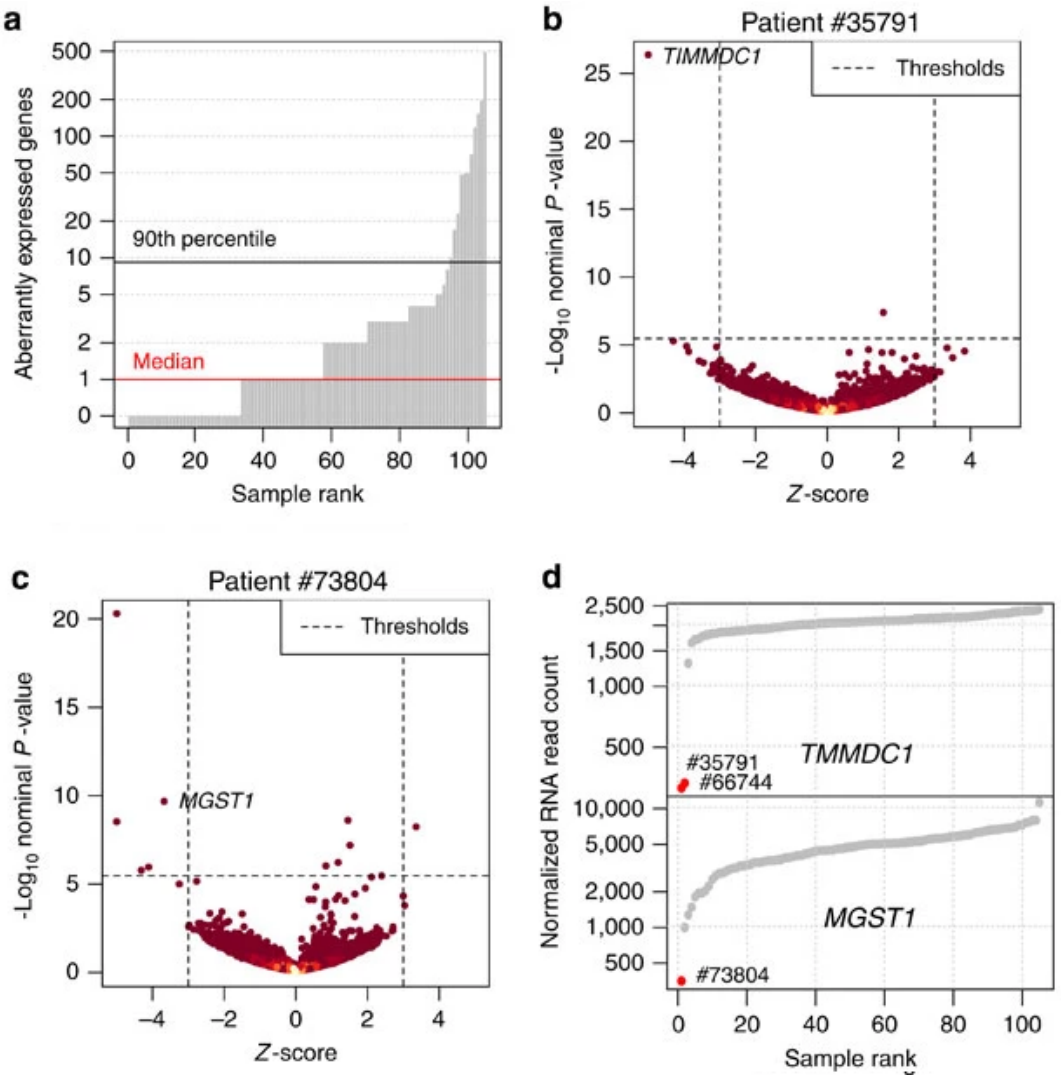

Figure 2.2: Aberrant expression detection in RNA-seq data. (a) Aberrantly expressed genes (Hochberg corrected P value < 0.05 and |Z-score| > 3) for each patient fibroblasts. **(b)** Gene-wise RNA expression volcano plot of nominal P values (-$\log_{10}$ P value) against Z-scores of the patient #35791 compared against all other fibroblasts. Z-scores with absolute value > 5 are plotted at ± 5, respectively. **(c)** Same as b for patient #73804. **(d)** Sample-wise RNA expression is ranked for the genes TIMMDC1 (top) and MGST1 (bottom). Samples with aberrant expression for the corresponding gene are highlighted in red (#35791, #66744, and #73804). Adapted from Kremer et al. [2017].

ing Hochberg's FWER method.[Hochberg, 1988] As LeafCutter is reporting the P value on the cluster level and does not perform any classification of the event, we inspected manually all significant splicing events (Hochberg adjusted P value < 0.05) of the 48 undiagnosed cases and classified them as exon skipping, exon truncation, exon elon-

gation, new exon, complex splicing (any other splicing event or a combination of the aforementioned ones), and false positives.

Applying our adaptation of LeafCutter for rare diseases resulted in a median of 5 abnormal events per sample, whereas 90% of samples had < 16 events and only 4 samples had > 100 events (Hochberg adjusted P value < 0.05, Fig. 2.3a). A total of 189 aberrantly spliced genes were detected in the 48 undiagnosed cases, with exon skipping and differential expression being the main cause of aberrant splicing followed by the creation of new exons, while LeafCutter incorrectly predicted aberrant splicing in 14 cases (Fig. 2.3b). Searching for genes encoding mitochondrial proteins, we found *CLPP* and *TIMMDC1* among the 20 most significant aberrantly spliced genes. In the case of *CLPP* #58955 showed an exon-skipping event, where 82 out of 136 split reads skipped exon 5, while 14 additional reads truncated exon 5 on the 3' end, resulting in $\psi_3 = 29.4\%$ wildtype *CLPP* (3' percent spliced in[Pervouchine et al., 2013]). The likely genetic cause of these two splice defects is a rare homozygous variant in exon 5 of *CLPP* affecting the last nucleotide of exon 5 (c.661G>A, p.Glu221Lys, 2.6×10^{-5} minor allele frequency (MAF) with only heterozygous variant states reported in the gnomAD database[Karczewski et al., 2020]). Both the exon skipping and exon truncation event result in truncated CLPP and western blots corroborated the complete loss of full-length CLPP (Kremer et al. [2017] Supplementary Fig. 5 and 11). Initially this variant was classified as VUS and was only as candidate gene reported among 30 other potentially bi-allelic candidate genes. Only with this additional information on a functional level through the splicing analysis and the confirmed loss of expression by western blotting provides enough evidence to reclassify this variant as disease causing. In top of this, *CLPP* encodes a mitochondrial ATP-dependent endopeptidase[Halperin et al., 2001] and CLPP-deficiency causes Perrault syndrome[Jenkinson et al., 2013, 2012] (OMIM #601119) which is overlapping with the clinical presentation of the patient investigated here including microcephaly, deafness and severe psychomotor retardation. Moreover, a study published around the same time showed that Clpp-/- mice are deficient for complex IV expression,[Szczepanowska et al., 2016] in line with complex IV deficiency of this patient.

As *TIMMDC1* was already found as aberrantly expressed gene, the splicing analysis provided additional information on the pathomechanism of this event. In both cases, the RNA-seq analysis revealed that primarily a *TIMMDC1*-isoform was expressed that included a new exon deep in intron 5 (Fig. 2.3c). This new exon introduces a frameshift yielding a premature stop codon (p.Gly199_Thr200ins5*, Fig. 2.3d). Moreover, this new exon contained a rare variant (c.596+2146A>G) not listed in the 1,000 Genomes Project[The 1000 Genomes Project Consortium, 2015] nor in gnomAD.[Karczewski et al., 2020] We used Sanger sequencing to validate the presence of the homozygous variant in each case. WGS did not identify any other rare variant in and around *TIMMDC1* but confirmed the new variant 6bp inside the new exon. Only 2 out of 6 splicing prediction tools predicted an impact of this variant on splicing.[Piva et al., 2012; Dogan et al., 2007; Timmermans et al., 2010; Desmet et al., 2009; Yeo et al., 2004b; Burge and Karlin, 1997] Specifically, SpliceAid2 predicted multiple binding sites for splice enhancers,[Piva et al., 2012] while SplicePort predicted the usage of the new acceptor and donor sites (feature

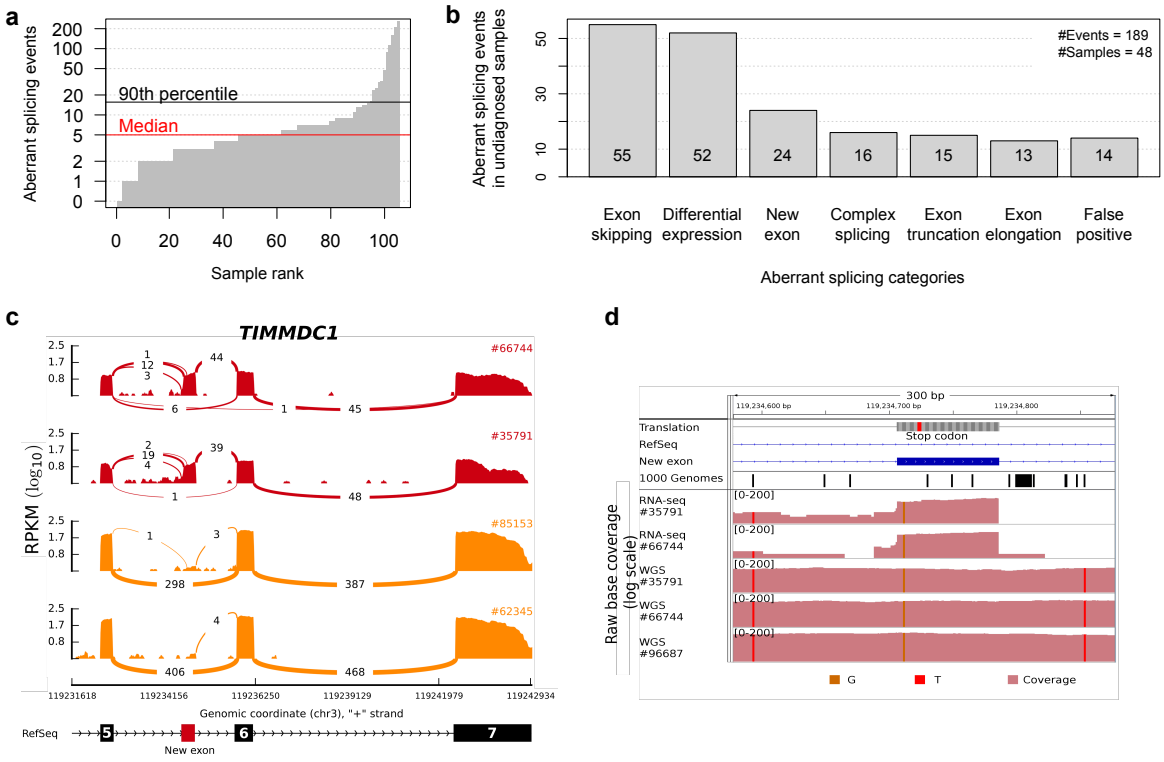

Figure 2.3: Aberrant splicing detection and quantification. **(a)** Aberrant splicing events (Hochberg corrected P value < 0.05) for all fibroblasts. **(b)** Aberrant splicing events ($n = 175$) in undiagnosed patients ($n = 48$) grouped by their splicing category after manual inspection. **(c)** TIMMDC1 sashimi plot of a cryptic exon creation event in TIMMDC1-affected and TIMMDC1-unaffected fibroblasts (red and orange, respectively). The RNA coverage is given as the $\log_{10}$ RPKM-value and the number of split reads spanning the given intron is indicated on the exon-connecting lines. At the bottom the gene model of the RefSeq annotation is depicted and the aberrant event is coloured in red. **(d)** Coverage tracks (light red) for patients #35791, #66744, and #91324 based on RNA and WGS. For patient #91324 only WGS is available. The homozygous SNV c.596+2146>4G is present in all coverage tracks (vertical orange bar). The top tracks show the genomic annotation: genomic position on chromosome 3, DNA sequence, amino acid translation (grey, stop codon in red), the RefSeq gene model (blue line), the predominant additional exon of TIMMDC1 (blue rectangle) and the SNV annotation of the 1000 Genomes Project (each black bar represents one variant). Adapted from [Kremer et al., 2017].

generation algorithm score 0.112 and 1.308, respectively). [Desmet et al., 2009] A reevaluation of our in-house WGS database revealed another case with the same homozygous variant. In this family three affected siblings presented with similar clinical symptoms although without a diagnosis of a mitochondrial disorder (Fig. 2.3d). Two siblings died before the age of 10 while the youngest brother (#96687) was still alive at age of 6. The

discovery of the same intronic *TIMMDC1* variant in three unrelated families from three different ethnic groups with similar clinical presentations together with the detected splicing defects in RNA-seq provide convincing evidence for the causality of this variant for TIMMDC1 loss of function.

2.2.3 Mono-allelic expression

To detect MAE, we started with the variant calls from WES data. We applied the same variant filters as described in Kremer et al. [2017]. In order to get a full picture of MAE, we selected only heterozygous SNVs and did not apply any variant frequency filter. We filtered out any SNV where multiple alleles at the same position was detected to reduce noisy calls. For all remaining SNVs, we used the *pileLettersAt* function from the R/Bioconductor package GenomicAlignments[Lawrence et al., 2013] to extract the RNA-seq coverage. After counting, we discarded any variant that was not covered by at least 10 reads. Then, we used the DESeq2 package[Love et al., 2014] to assess the significance of the allele-specific expression. To this end, we compared within each sample the coverage of the wildtype allele with the coverage of the alternative allele. For simplicity, we treated each SNV as independent event and ignored the fact that some variants could be in linkage. Specifically, we used a generalized linear model per sample to fit k_{ij}, the number of reads of the variant j in condition $i \in \{\text{wildtype}, \text{alternative}\}$, as:

$$\begin{aligned} k_{ij} &\sim \text{NB}(s_i \times q_{ij}, \alpha) \\ \log_2(q_{ij}) &= \beta_j^0 + \beta_j^{\text{allele}} x_{ij}^{\text{allele}}, \end{aligned} \tag{2.2}$$

where NB is the negative binomial distribution and s_i is the size factor of the condition. We set a global dispersion parameter to $\alpha = 0.05$, which resembles the average dispersion value based on the aberrant expression analysis (Section 2.2.1). β_j^0 is the intercept parameter for variant j, while x_{ij}^{allele} was set to 1 for the alternative allele and 0 for the wildtype allele. The resulting β_j^{allele} represents the $\log_2$-fold changes for the alternative allele against the reference allele. Due to our DESeq2 design, we disabled the independent filtering (*independentFiltering = FALSE*) to keep all variants in the result table independent of their coverage. Each variant was classified as mono-allelically expressed if $|\beta_j^{\text{allele}}| \geq 2$, which corresponds to an allele frequency ≥ 0.8, and was said to be significant if it had a multiple testing corrected P value < 0.05. We controlled for the false discovery rate (FDR) per sample across all variants with the Benjamini-Hochberg method.[Benjamini and Hochberg, 1995]

Screening for MAE in the 85 samples with matching WES and RNA-seq, we found per sample in median 34,064 heterozygous SNVs detected by WES. In median, 7,529 of them passed our RNA-seq coverage filter (coverage ≥ 10), while only 20 showed MAE (FDR adjusted P value < 0.05 and allele frequency ≥ 0.8), of which 6 were rare variants (MAF < 0.001 (Fig. 2.4a). The MAE analysis 610 events in total for 85 samples, but did not reveal any extreme outlier sample as happened in the aberrant expression and splicing analysis; 25 of rare MAE events in one sample was the maximum. Amongst the 18 rare MAE events in patient #80256, the VUS in *ALDH18A1* caught our attention

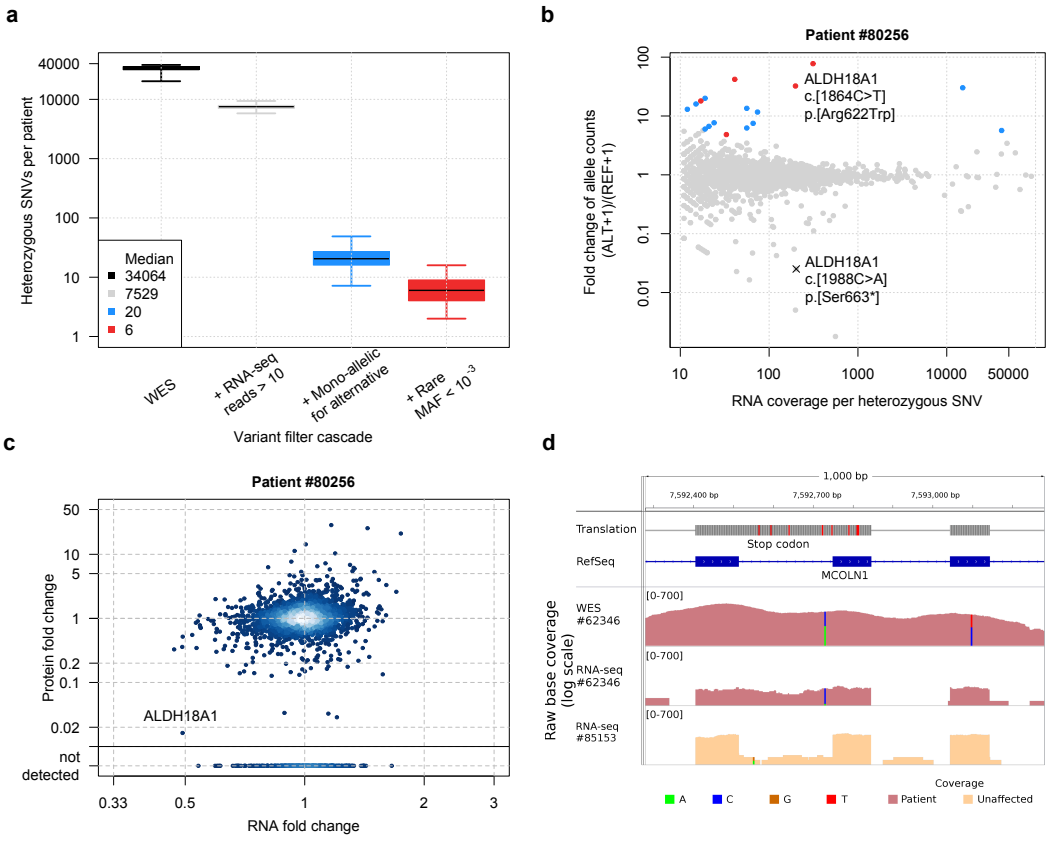

Figure 2.4: Detection of mono-allelic expression of rare variants. **(a)** Distribution of heterozygous SNVs across samples for different consecutive filtering steps. Heterozygous SNVs detected by WES (black), SNVs with RNA-seq coverage of ≥ 10 reads (grey), SNVs with an alternative allele frequency > 0.8 and a Benjamini-Hochberg corrected P value < 0.05, blue), and subsetted to rare SNVs (ExAC MAF < 0.001, red). **(b)** Fold change between alternative (ALT+1) and reference (REF+1) allele read counts for the patient #80256 compared to total read counts per SNV within the sample. Points are coloured according to the groups defined in a. **(c)** RNA fold changes plotted against protein fold changes for case #80256. The position of *ALDH18A1* is highlighted. Reliably detected proteins that were not detected in this sample are shown separately with their corresponding RNA fold changes (points below solid horizontal line). **(d)** Intron retention for *MCOLN1* in patient #62346. Tracks from top to bottom: genomic position on chromosome 19, amino acid translation (red for stop codons), RefSeq gene model, coverage of WES of patient #62346, RNA-seq based coverage for patients #62346 and #85153 (red and orange shading, respectively). SNVs are indicated by non-reference coloured bars with respect to the corresponding reference and alternative nucleotide. Adapted from Kremer et al. [2017].

(c.1864C>T, p.Arg622Trp, Fig. 2.4b). *ALDH18A1* encodes an enzyme involved in mitochondrial proline metabolism.[Adams and Frank, 1980] This particular VUS had been

picked up by WES in conjunction with a nonsense variant (c.1988C>A, p.Ser663*, Fig. 2.4b). Because at the time of the WES analysis *ALDH18A1* was associated only with curtis laxa III (OMIM #138250),[Baumgartner, 2000; Fischer-Zirnsak et al., 2015] which the patient did not present, the compound heterozygous variants were not followed up. Due to the re-prioritization through the MAE analysis, we investigated the causality of the *ALDH18A1* variants. Quantitative proteomics showed almost complete loss of functional ALDH18A1 ($\sim$ 2% of normal ALDH18A1 protein abundance, Fig. 2.4c) highlighting the potential effect of the rare VUS on the protein abundance. Metabolomics profile of blood plasma was in accordance with a defect in proline metabolism (Kremer et al. [2017] Fig. 4d) and the following changes in urea cycle. In addition, we were able to rescued the growth rate of the fibroblasts through supplementation of proline, linking the impaired proline metabolism to the detected MAE variants. Finally, in another study, ALDH18A1-deficiency was linked to spastic paraplegia without cutis laxa (OMIM #138250),[Coutelier et al., 2015] matching the patient's clinical presentations validating these ALDH18A1 variants as disease-causing.

While lowering the filtering thresholds and combining the information from the three analysis to find even more candidates, we noticed *MCOLN1* in case #62346. As the most down-regulated gene in case #62346, it was not prioritized with our stringent cutoffs despite its reduced expression level ($p_{\text{adj}} = 0.065$ and z score $= -2.97$). In addition, the MAE analysis revealed two rare variants in a true compound heterozygous state, one expressing the reference allele (12 REF vs 1 ALT reads, c.832C4T, p.Gln278*) and the other expressing the alternative allele (1 REF vs 10 ALT reads, c.681-19A>C) indicating that only one allele is expressed. Due to the low coverage, both events did not reach significance but were prioritized due to their extreme allele frequencies (0.08 and 0.91, respectively). The loss of expression of one allele is probably due to nonsense-mediated decay as response to the nonsense variant. Further investigation of the intronic variant showed that it was part of an intron retention that introduced a nonsense codon (p. Lys227_Leu228ins16*, Fig. 2.4d). The compound heterozygous variants were initially missed by WES analysis because the intronic variant was classified as VUS, although mucolipidosis (OMIM #605248),[Sun, 2000] an associated phenotype of *MCOLN1*, was matching the patient's clinical presentation. Also additional enzymatic tests available at the time for mucolipidosis types I, II, and III did not show any enzyme deficiency in blood leukocytes. In contrast to WES, RNA-seq enabled the detection of the two loss-of-function alleles in *MCOLN1* and therefore established the genetic diagnosis in patient #62346.

2.3 Weak splicing

It is already known that cryptic splice sites are actively expressed at low levels and that they can be activated by a single point mutation.[Kapustin et al., 2011] This is inline with our observation in all non-TIMMDC1 deficiency samples as they expressed in low levels the intron-junction to and from the new cryptic exon that is activated by a SNV (Fig. 2.3c). To systematically assess this phenomenon of cryptic splice sites and

their linkage to cryptic exons in a genome-wide fashion, we modelled all intron-centric ψ (percent spliced in) values [Pervouchine et al., 2013]. To this end, we only considered reads spanning an exon-exon junction, so called split reads, with a mapping quality of > 10 to reduce the false-positive rate due to mapping issues. The intron-centric ψ_5 and ψ_3 values were computed according to Eq. 1.1 and 1.2 as described by Pervouchine et al. [2013] across all samples for a given dataset. By modelling the distribution of the ψ values with three components, we classified the splice sites into: *background*, *weak*, and *strong*. To link the introns to a specific category, we used the annotation status provided by GENCODE.[Harrow et al., 2012] We used GENCODE release 24 to group the introns into: (i) *both sites annotated* where both splice sites of the given intron are present in GENCODE, (ii) *one site annotated* where only on of the sites are annotated in GENCODE, and (iii) *no side annotated* where neither of the splice sites are annotated in GENCODE. We specifically modelled the number of split reads $\mathrm{n}(D, A)$ of a intron conditioned on the total number of reads $\mathrm{N}(D, A) = \sum_{A'} \mathrm{n}(D, A')$ for ψ_5 and $\mathrm{N}(D, A) = \sum_{D'} \mathrm{n}(D', A)$ for ψ_3 as:

$$P\left(\mathrm{n}(D, A)|\mathrm{N}(D, A)\right) = \sum_{c\in\{bg,wk,st\}} \sum_{s\in\{0,1,2\}} \pi_{c,s}\mathrm{BB}\left(\mathrm{n}(D, A)|\mathrm{N}(D, A), \alpha_c, \beta_c\right), \tag{2.3}$$

where c is the component index, s the number of annotated sites (group index) and BB the beta-binomial distribution. Hence, the components were modelled to have the same parameters α_c and β_c in all three groups but their mixing proportions $\pi_{c,s}$ to be group-specific. Fitting was performed using the expectation-maximization (EM) algorithm. For the initial step, the data points were classified as *background* ($\psi \leq 0.001$), weak spliced ($0.001 < \psi \leq 0.1$) and canonical ($\psi > 0.1$). After convergence of the clustering the obtained parameters were used to estimate the probability for each side of the intron to belong to a given class. In order to validate our findings from the Kremer dataset, we applied the same algorithm to each tissue of the genotype-tissue expression (GTEx) dataset (V7).[The GTEx Consortium et al., 2015] As the GTEx dataset includes more than 12,000 postmortem RNA-seq samples from 714 assumed healthy donors from different origins over 53 different tissues, it servers as a good reference for the expected expression levels. A more detailed summary of the GTEx dataset can be found in Section 3.2.

In the Kremer dataset, we computed genome-wide the ψ_5 and ψ_3 values for the 1,603,042 observed splicing events in 119 RNA-seq samples after quality filtering. Modelling the ψ distribution resulted in the classification of splicing events to be *strong*, *weak*, and *background* in 20%, 16%, and 64% of the events with the ψ boundaries $\psi > 5.29\%$, $0.16\% < \psi \leq 5.29\%$, and $< 0.16\%$, respectively (Fig. 2.5 and A.5a-b). Looking at the average ln-likelihood of our fit confirms that our EM algorithm converged within 250 iterations (Fig. 2.5c). Using these classifications, we looked at all detected private exons and discovered that 17 out of the 24 events (70% which is 4.4-fold more than by chance) originated from weak splice sites (Fig. 2.5 bottom). To confirm these results in a healthy cohort, we used the GTEx dataset and run the same analysis. The global ψ distribution shifted for all groups slightly towards higher ψ values compared to the Kremer dataset

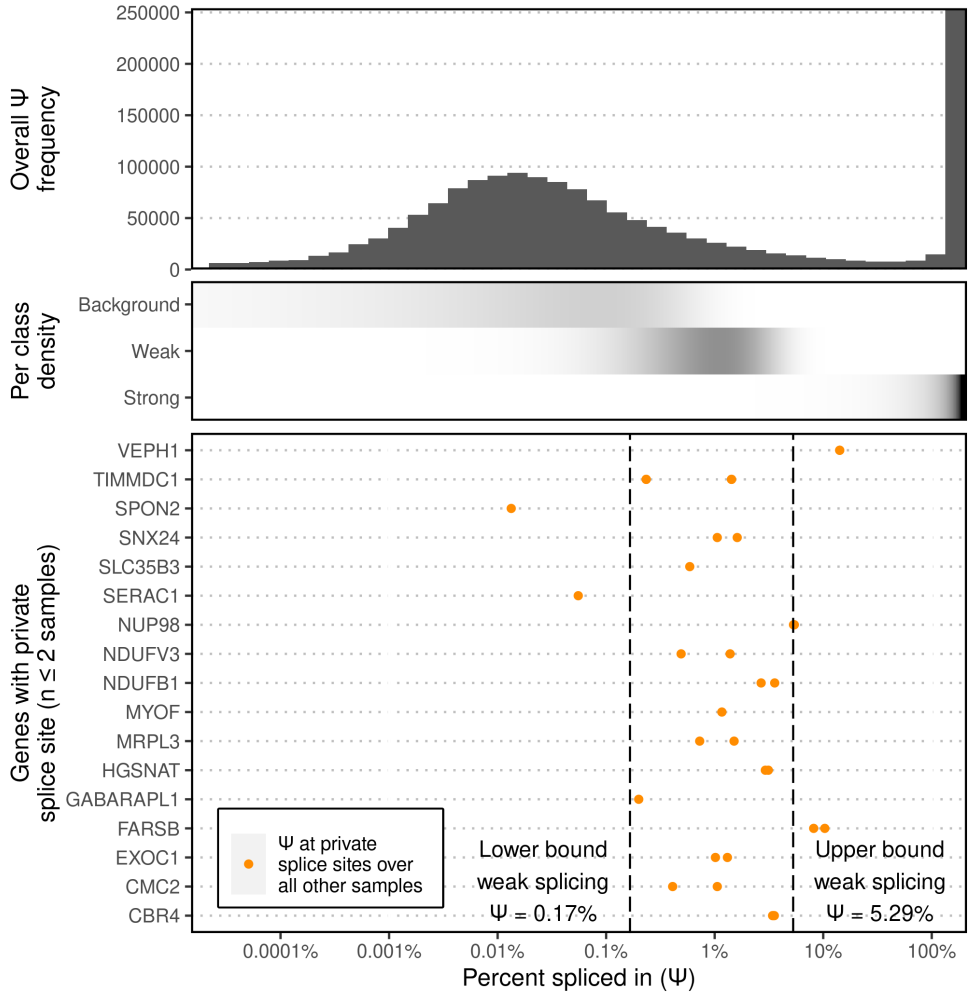

Figure 2.5: Weak splicing leading to cryptic exons. Percent spliced in (Ψ) distribution for different splicing classes and genes. Top: histogram of the genome-wide distribution of both ψ_5 and ψ_3 values based on all reads over all samples. Middle: The shaded horizontal bars represent the densities (black for high density) of the background, weak and strong splicing class, respectively. Bottom: ψ values of the predominant donor and acceptor splice sites connecting cryptic exons (aberrantly expressed in at most two samples) computed over all other samples. The dashed lines represent the lower and upper boundaries for the weak splicing class. Adapted from Kremer et al. [2017].

(Fig. 2.6a-c). The most notable change was in the *no side annotated* group from an unimodal towards bimodal distribution with two peaks around 0.1% and 100% ψ, respectively. This means that in GTEx overall more novel introns were detected compared to the Kremer dataset. Fitting the model revealed that the geometric mean across all GTEx tissues for the lower boundary of the weak splicing class with $\psi = 0.22\%$ did not change much from the $\psi = 0.16\%$ in the Kremer dataset (Fig A.6). But the upper boundary did change from $\psi = 5.29\%$ to $\psi = 52.21\%$, which is probably due to the shift

of the global ψ distribution. Despite the change of the upper boundary, 19 out of the 24 (79%) detected cryptic splicing events were still classified as a *weak* splice site (Fig. 2.6b). These findings with the Kremer and GTEx dataset confirm that weakly spliced cryptic exons are loci more susceptible to become strongly spliced sites over other deep intronic regions. Often, these weak splice sites require only a single nucleotide variant to activate the site and incorporate the cryptic exon into the major isoform, which can have tremendous impact on the phenotype like the full down-regulation of TIMMDC1. In standard RNA-seq analysis pipelines only the major isoforms and annotated introns are considered, while such weakly spliced introns are removed as noisy data. But our analysis show that in the case of rare disease diagnostics, the annotation of weak splice sites through accumulation of reads across multiple samples can help to prioritize deep intronic VUSs detected by WGS.

2.4 Short summary

We demonstrated the power of RNA-seq to support molecular diagnostics in rare diseases in three ways: (i) discovery of a new disease-associated gene, (ii) diagnosis of 10% (5 of 48) of undiagnosed cases, and (iii) identification of a limited number of strong candidates. This was achieved by combining the high-throughput RNA-seq technology with newly developed outlier detection algorithms able to detect aberrant gene expression, aberrant splicing events, and mono-allelic expression of the rare variants. Our significance-based algorithms detected in median 1, 5, and 6 outlier events per sample, respectively, a manageable size of candidate genes for manual inspection and validation. Overall, we identified strong candidate genes in known disease-causing or mitochondrial protein-coding genes in 36 of 48 cases. Through the identification of aberrant expression events, classification of weak splice sites, and the reclassification of multiple VUS, we highlight the benefits of using RNA-seq to improve the interpretation of variants for rare diseases but also in general. Overall, we proved the relevance and complementarity of RNA-seq for rare disease diagnostics. This study also revealed the need for specialized algorithms to detect aberrant events in RNA-seq data.

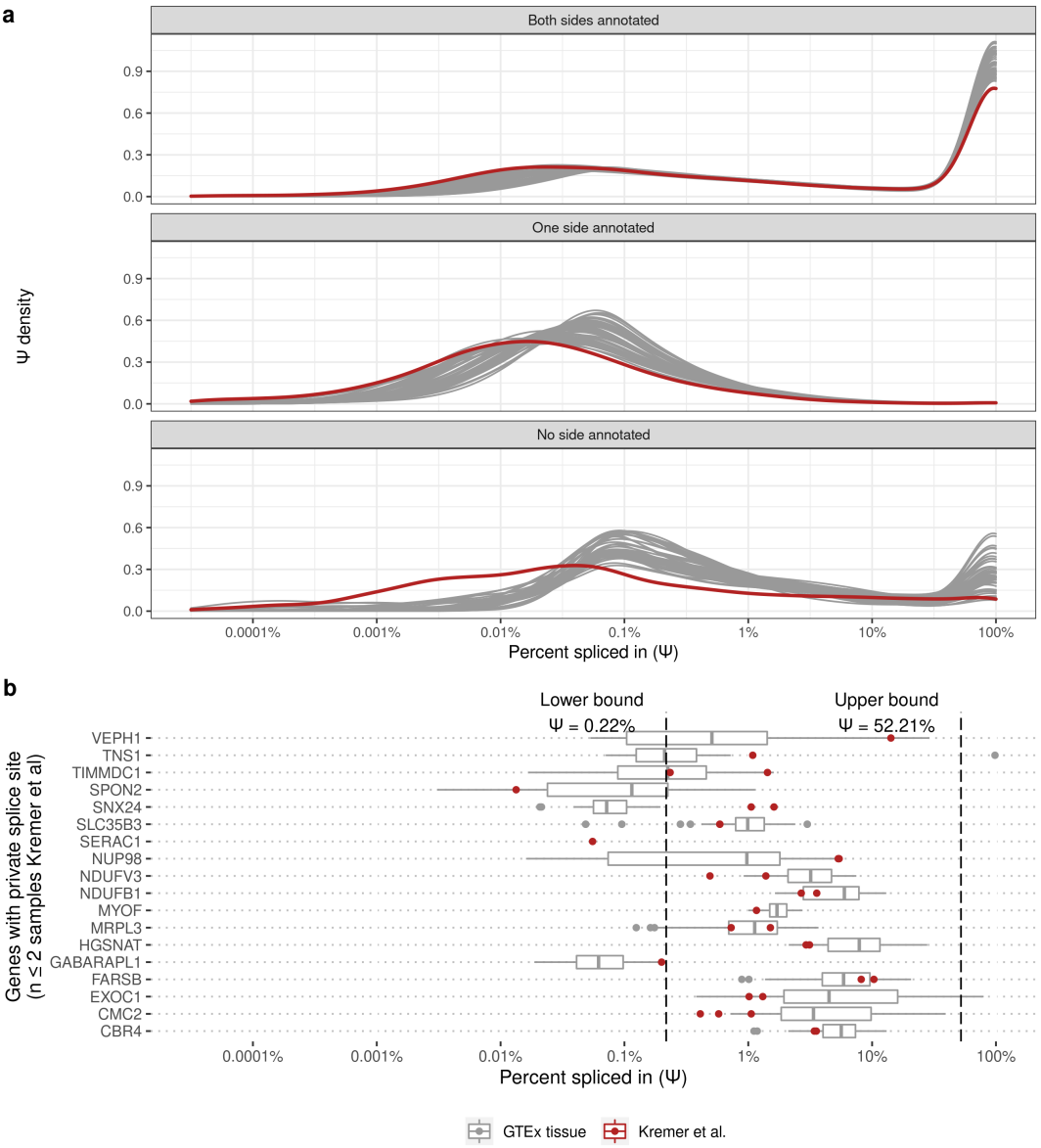

Figure 2.6: Weak splicing in GTEx tissues. **(a)** Density (y-axis) of the genome-wide distribution of both ψ_5 and ψ_3 values (x-axis) for exon-exon junctions based on all reads over all samples per GTEx tissue (gray lines). The red line shows the distribution presented by Kremer et al. [2017]. The data is stratified by the exon-exon junction's annotation status based on GENCODE[Harrow et al., 2012]: (i) both ends are present in GENCODE, only one end is present in GENCODE, neither ends are present in GENCODE. **(b)** ψ value distribution across GTEx tissues for exon-exon junctions leading to aberrantly expressed cryptic exons in Kremer et al. [2017]. The ψ values are computed on all reads over all samples per tissue. The red points depict the ψ value observed by Kremer et al. [2017] across the non-effected samples. The dashed line depicts the lower and upper boundary for the weak splicing class averaged across all GTEx tissues.

3 Detection of aberrant gene expression with OUTRIDER

The methodology, results, and figures presented in this chapter are part of the manuscript "OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data" from Brechtmann et al. [2018]. The author's contributions are included in it's preprint on bioRχiv[1]. In short, I developed the software together with Felix Brechtmann and Agnė Matusevičiūtė. I critically analysed and visualized the data together with the other co-authors. The manuscript was written by Felix Brechtmann, me, and Julien Gagneur with the help of Vicente Yépez and Agnė Matusevičiūtė.

3.1 Motivation

In our pilot study (Section 2), we were able to increase the diagnostic rate by 10% through the detection of aberrant events in RNA-seq data.[Kremer et al., 2017] In parallel, Cummings et al. [2017] also investigated the potential of using RNA-seq in diagnostic and confirmed that RNA-seq is complementary to WES and WGS by providing functional interpretation of regulatory defects. Cummings et al. [2017] even improved the diagnostic rate by 35% over WES/WGS by revealing splicing defects, mono-allelic expression of heterozygous loss-of-function variants, and gene expression outliers. As no sophisticated method was available to detect aberrant events in RNA-seq data at the time, the two studies differed in their approaches despite the same goals. Cummings et al. [2017] detected expression outliers by computing z scores on log-transformed gene-length-normalized read counts by subtracting the mean count and dividing by the standard deviation. Then a z score cutoff of $|zscore| > 3$ was used to identify expression outliers without a formal statistical assessment of the events. This approach did not reveal any convincing pathogenic expression outlier candidates. Cummings et al. [2017] concluded that the cohort was too small and hence underpowered to detect any expression outliers. In contrast, we were able to find 4 out of 6 disease-causing aberrant events as expression outliers in our pilot study (Section 2.2.1) even with a smaller cohort size ($n = 119$ versus $n = 230$ samples in Kremer et al. [2017] and Cummings et al. [2017], respectively). As described in Section 2.2.1, we applied a stringent significance test together with the z score cutoff approach (FWER adjusted $p < 0.05$ and $|zscore| > 3$). We used DESeq2[Love et al., 2014], a method developed for differential expression analysis,

[1]Brechtmann, F. et al. OUTRIDER: A statistical method for detecting aberrantly expressed genes in RNA sequencing data. bioRχiv 322149; doi: 10.1101/322149

to test each sample against the rest of the cohort using a negative binomial (NB) distribution to assess the significance. Based on the results of the two studies, it remained unclear what caused the difference. Some of the uncertainty can be attributed to the small number of individuals diagnosed, the lack of ground truth, a direct comparison of methods, and adequate benchmarking of the methods themselves.

Another big difference between the two studies is how biases and confounders in the RNA-seq data were controlled for. Cummings et al. [2017] only controlled for variation in sequencing depth by using reads per kilobase per million mapped reads (RPKM) values in the z score calculation while ignoring other potential confounders. As age, genetic factors, sex, RNA integrity number (RIN), technical biases, and other known and unknown confounders can influence the downstream analysis, it was shown in multiple studies that accounting for them and controlling the read counts accordingly improves the outcome.[Leek and Storey, 2007; Dillies et al., 2013; Stegle et al., 2012] For example, down-regulation of a Y chromosome-encoded gene in males would not be detectable unless one would control for the sex of the samples. Therefore, we manually controlled for sex, biopsy site, and technical variations inferred from hierarchical clustering in our study (Section 2.2.1). In another study that also detected expression outliers by computing z scores, Li et al. [2017] controlled for sex, the top three genotype principal components, as well as for hidden confounding effects estimated by the probabilistic estimation of expression residuals (probabilistic estimation of expression residuals (PEER)) method.[Stegle et al., 2012]. While all three studies controlled in one way or another for covariations in the RNA-seq read count data, none of them assessed or tuned it for the detection of aberrantly expressed genes.

Therefore, we developed OUTRIDER (Outlier in RNA-Seq Finder), an algorithm that provides a statistical test for outlier detection in RNA-seq samples while controlling for covariations among the gene read counts (Fig. 3.1). OUTRIDER uses a denoising autoencoder to automatically model known and unknown confounders and assess the significance with the NB distribution allowing for overdispersed RNA-seq read count data. Further, we evaluated the added value of each of the two components towards expression outlier detection accuracy and developed and applied a benchmark strategy to compare OUTRIDER against the state-of-the-art methods at the time utilizing simulated data and the two experimental datasets from Kremer et al. [2017] and the GTEx consortium.[The GTEx Consortium et al., 2015]

3.2 Dataset description

To develop, test, and benchmark the OUTRIDER approach, we considered 4 different datasets covering different scenarios. The first dataset is composed of the RNA-seq samples of the rare disease cohort used in the pilot study described in Section 2.2. We used this dataset to benchmark the methods in a rare disease diagnostic setting. For the second experimental dataset we used the RNA-seq samples from the GTEx project.[GTEx Consortium, 2017]. All samples are collected postmortem but the donors are assumed healthy with no underlying condition and were not under treatment. Nevertheless,

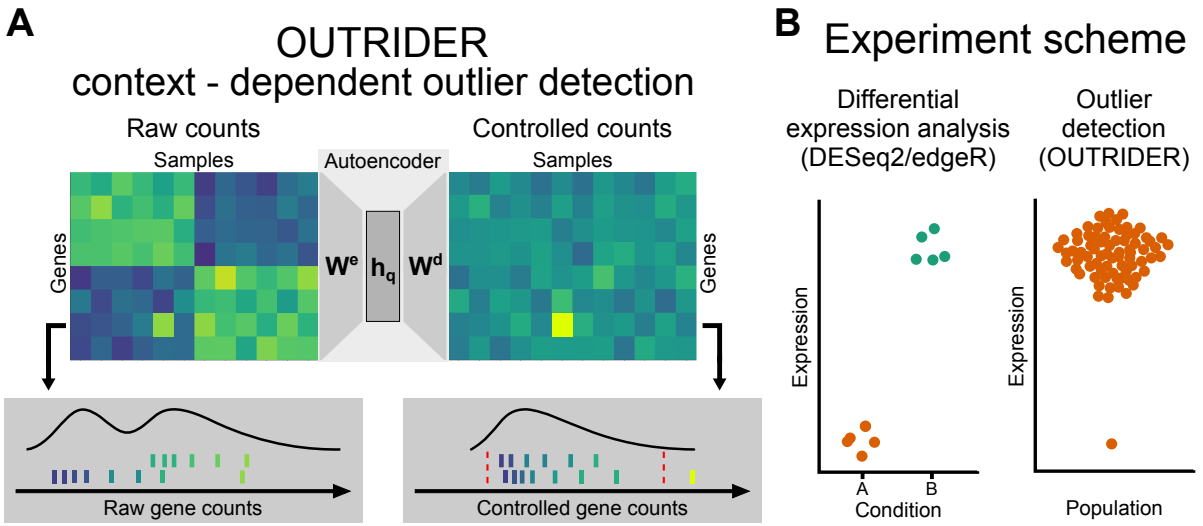

Figure 3.1: OUTRIDER overview (A) Context-dependent outlier detection. The algorithm identifies gene expression outliers whose read counts are significantly aberrant given the covariations typically observed across genes in an RNA-seq dataset. This is illustrated by a read count (left panel, fifth column, second row from the bottom) that is exceptionally high in the context of correlated samples (left six samples) but not in absolute terms for this given gene. To capture commonly seen biological and technical contexts, an autoencoder models covariations in an unsupervised fashion and predicts read-count expectations. Comparing the earlier mentioned read count with these context-dependent expectations reveals that it is exceptionally high (right panel). The lower panels illustrate the distribution of read counts before and after controlling for covariations for the relevant gene. The red dotted lines depict significance cutoffs. **(B)** Schema showing the differences in the experimental designs for differential expression analyses and outlier detection analyses; relevant analysis packages are mentioned. Taken from Brechtmann et al. [2018].

aberrant gene expression in these samples has been reported.[Li et al., 2017] The GTEx dataset, as we will call it from now on, will serve as the healthy control benchmark set. To have comparable results, we focused if not stated differently on the 250 suprapubic skin samples as they are the closest to the fibroblast cell lines from the Kremer dataset. The gene read counts were downloaded from the GTEx portal (V6P, counted with RNA-SeQC v1.1.8[DeLuca et al., 2012]) and only samples with a RIN $\geq$ 5.7 were considered. Genes were considered expressed and kept for downstream analysis if at least 5% of the samples had a fragments per kilobase per millions of reads (FPKM) value greater than 1 and more than 25% of the samples at least 1 read (Fig A.2d and 3.2a). The FPKM values were calculated with DESeq2,[Love et al., 2014] where the gene length was defined as the aggregated length of all exons. This resulted in 119 and 249 samples and 10,556 and 17,065 genes for the Kremer and GTEx dataset, respectively. The third and fourth dataset was simulated by drawing from a NB and log-normal distribution, respectively.

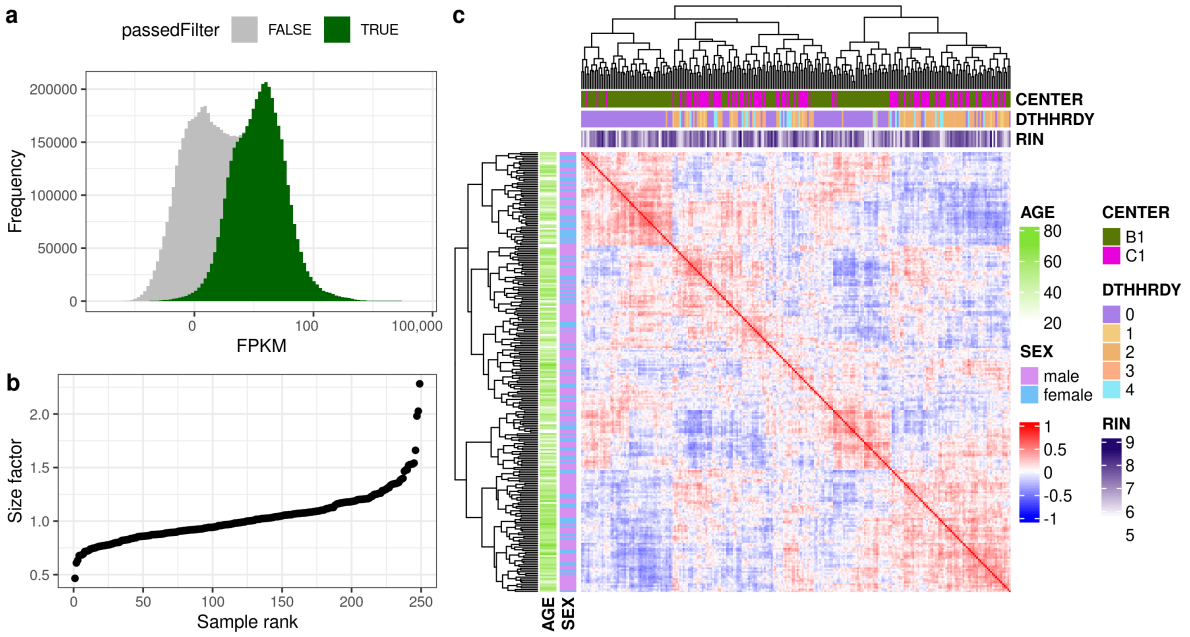

Figure 3.2: Dataset overview of the GTEx suprapubic skin tissue. **(a)** Histogram of the FPKM values grouped according to the filter status. Green indicates the genes that passed the filter and gray those that were filtered out. **(b)** Size factor plotted against the rank. Each dot represents a sample. **(c)** Correlation matrix of row-centered log-transformed read counts (249 samples and 17,065 genes). Red indicates a positive correlation and blue a negative correlation. The dendrogram represents the sample-wise hierarchical clustering. Colored horizontal and vertical tracks display sequencing center, cause of death (DTHHRDY, Hardy scale classification), RNA integrity number (RIN), gender, and age of the samples.

Before drawing the read counts, we fixed the latent space to have ten dimension. In the case of the log-normal distribution, we rounded the final number to the closest integer.

3.3 Statistical modeling

Based on the knowledge from previous studies,[Kremer et al., 2017; Cummings et al., 2017] we aimed to tackle two major limitations. The first one was to control the RNA-seq data for known and unknown confounders as they, if not removed, can dramatically lower the statistical power. The second limitation was that no specialized method was developed to assess the statistical significance of expression outlier events in RNA-seq data.

To control for known and unknown covariation in the read count data, we applied a normal and denoising autoencoder schema. Autoencoders are used to find a representation, also called encoding, in a lower dimension and so are typically used in dimensionality reduction in high-dimensional data in an unsupervised manner.[Lecun, 1987; Bourlard and Kamp, 1988; Hinton and Zemel, 1994] They have been shown to

be useful for extracting meaningful biological features as well as for imputing missing values in bulk and single cell RNA-seq data.[Way and Greene, 2018; Eraslan et al., 2019; Lopez et al., 2018; Kinalis et al., 2019] Denoising autoencoders are a subclass of autoencoders where the input is reconstructed while being corrupted by typically zeroing out up to 50% of the data points or adding noise.[Vincent et al., 2008, 2010] While we used a denoising autoencoder to optimize the hyperparamter q, the dimension of the latent space, by recalling corrupted read counts, a normal autoencoder was used to control the common covariation patterns among genes by learning the latent space of the input data (Fig. 3.1a).

In a rare disease diagnostic setting, the experimental design differs from the well established differential expression analysis workflows. In rare diseases, every case has its own individual genetic cause of disease even though the resulting phenotype can be similar. Also replicates are often not available. Hence, a typical case versus control comparison as done in differential expression analysis[Love et al., 2014; Zhou et al., 2014] can not be applied. It is more common to have several samples up to hundreds that serve as the population and each sample is then tested per gene if it deviates from its population (Fig. 3.1b). It is noteworthy that DESeq2[Love et al., 2014] and edgeR[Zhou et al., 2014] already have procedures to mark or downweight outlier points using Cook's distance and Pearson residuals, respectively. But the purpose of these procedures are to increase the robustness of the model fit by removing outliers instead of assessing the significance of them, which is the aim in rare disease diagnostics. This ultimately leads to outlier detection for the univariate case, where the distribution of the population is modeled jointly while each data point is subsequently tested to assess whether it deviates significantly from the fitted distribution (Fig 3.1b).

We assume that the count k_{ij} of gene $j = 1, \ldots, p$ in sample $i = 1, \ldots, N$ follows a NB distribution accounting for overdispersed count data.[Whitaker, 1914] Specifically,

$$\begin{aligned} \mathrm{P}\left(k_{ij}\right) &= \mathrm{NB}\left(k_{ij} | \mu_{ij} = c_{ij}, \theta_j\right) \\ \mathrm{NB}\left(k | \mu, \theta\right) &= \frac{\Gamma(k+\theta)}{\Gamma(\theta) k!} \left(\frac{\mu}{\mu+\theta}\right)^k \left(\frac{\theta}{\mu+\theta}\right)^\theta, \end{aligned} \tag{3.1}$$

where θ_j is a gene-specific dispersion parameter and c_{ij} the expected count. The variance of the NB distribution is given by $\mathrm{Var} = \mu + \frac{\mu^2}{\theta}$. To prevent convergence issues in the lower range and overfitting in the upper range, we limited θ_j to the interval [0.01, 1000]. The expected count c_{ij} is the product of the sample-specific size factor s_i and the exponential of the factor y_{ij}:

$$c_{ij} = s_i \cdot \exp(y_{ij}). \tag{3.2}$$

We use the size factor s_i to control for technical variations in sequencing depth. The size factors are robustly estimated with the *estimateSizeFactor* function implemented in DESeq2.[Love et al., 2014]. The factors y_{ij} capture covariations across genes and are

modeled with an autoencoder of encoding dimension $1 < q < \min(p, N)$. Specifically,

$$\mathbf{y}_i = \mathbf{h}_i \mathbf{W}_d + \mathbf{b}, \tag{3.3}$$
$$\mathbf{h}_i = \widetilde{\mathbf{x}}_i \mathbf{W}_e, \tag{3.4}$$

where the $p \times q$ matrix $\mathbf{W}_e$ is the encoding matrix, the $q \times p$ matrix $\mathbf{W}_d$ is the decoding matrix, the q-vector $\mathbf{h}_i$ is the encoded representation, and the p-vector $\mathbf{b}$ is a bias term. Having a decoding matrix that is not the transpose of the encoding matrix, unlike for principal-component analysis (PCA), turned out to be important, most likely because the property that the matrix inverse equals the matrix transpose does not generalize to the NB loss function. The input vector to the autoencoder $\widetilde{\mathbf{x}}_i$ is centered by gene and computed as follows:

$$\widetilde{x}_{ij} = x_{ij} - \bar{x}_j, \tag{3.5}$$
$$\bar{x}_j = \text{mean}_i \left(x_{ij} \right), \tag{3.6}$$
$$x_{ij} = \log \left(\frac{k_{ij} + 1}{s_i} \right), \tag{3.7}$$

where we add 1 to prevent computing the logarithm of 0 and control for the sequencing depth by dividing by s_i. The combination of equations 3.2-3.7 is what we call the autoencoder implementation in OUTRIDER, in short, $c_{ij} = \text{AE}_{\text{OUTRIDER}}(k_{ij})$.

Fitting the autoencoder and negative-binomial distribution parameters Fitting the autoencoder is implemented as an iterative three-step procedure in which the parameters $\mathbf{W}_e$, $\mathbf{W}_d$, and q_j are iteratively updated until convergence. The autoencoder is initialized (i) by setting the encoder and decoder matrices to the loadings of a PCA using the *pca* function from the package pcaMethods,[Wold, 1966; Oba et al., 2003; Troyanskaya et al., 2001] (ii) by setting the bias vector to the mean of x_{ij} per gene (Eq. 3.6), (iii) by estimating the dispersion θ_j with the method of moments, (iv) an initial fit of all gene-specific parameters including the decoder matrix $\mathbf{W}_d$ and the dispersion θ_j. The autoencoder is then fitted through repetition of the following three update steps: (i) the encoder matrix is updated, (ii) the decoder matrix is updated, and (iii) the dispersion parameters are refitted. The steps ii and iii are parallelized over the genes. In each update step, the average negative log-likelihood is minimized with respect to the current parameters by the optimization method L-BFGS as implemented in *optim*.[Byrd et al., 1995; Team, 2021] Detailed derivations of the used loss functions and the respective gradients can be found in the Supplemental Data of Brechtmann et al. [2018]. The fitting procedure is stopped when the average negative log-likelihood of each step in one iteration does not differ more than the convergence threshold of 10^{-5} from the last step of the previous iteration or the maximum of 15 iterations is reached.

Fitting the optimal encoding dimension To find the optimal hyperparameter q, the latent space dimension, we applied a denoising autoencoder scheme. To this end, we

injected with a frequency of 10^{-2} corrupted counts k_{ij}^c into the read count data and measured the performance of the autoencoder to recall them. We computed the corrupted counts as follows:

$$u_{ij} = \log_2\left(\frac{k_{ij}}{s_i} + 1\right), \tag{3.8}$$

$$k_{ij}^c = \text{round}\left(s_i \cdot 2^{u_{ij} \pm e^z \sigma(u_j)}\right), \tag{3.9}$$

where $\sigma(u_j)$ is the standard deviation of the size factor normalized counts of gene j in the log space and z, the amplitude of the corrupted count. z is drawn from a normal distribution characterized by a mean of $\log(3)$ and a standard deviation of $\log(1.6)$. The sign of the shift is randomly selected. The optimal dimension q is then selected as the dimension maximizing the area under the precision-recall curve for identifying corrupted counts.

P value computation For every pair of gene j and sample i, we test the null hypothesis that the count k_{ij} follows a NB distribution as described by Equation 3.1. To detect down and up regulation of genes, we compute two-sided P values as:

$$P_{ij} = 2 \cdot \min\left\{\frac{1}{2}, \sum_{k=0}^{k_{ij}} \text{NB}(k_{ij}|c_{ij}, \theta_j), 1 - \sum_{k=0}^{k_{ij}-1} \text{NB}(k_{ij}|c_{ij}, \theta_j)\right\}. \tag{3.10}$$

Due to the nature of the discrete NB distribution, both one-sided P values can exceed $1/2$ at the same time, for which we introduced the $1/2$ term.

As we test per sample all expressed genes at the same time, we correct the P values for multiple testing using the FDR method. We use the Benjamini-Yekutieli procedure because it applies under positive dependence,[Benjamini and Yekutieli, 2001] which is necessary because gene expression is highly regulated and therefore genes in the same sample can be correlated, even after controlling for confounding effects by the autoencoder.

Z score computation Z scores z_{ij} are computed on a logarithmic scale as follows:

$$z_{ij} = \frac{l_{ij} - \mu_j^l}{\sigma_j^l}, \tag{3.11}$$

$$l_{ij} = \log_2((k_{ij} + 1)/(c_{ij} + 1)) \tag{3.12}$$

where l_{ij} is the log-transformed controlled count and μ_j^l and σ_j^l the mean and standard deviation of l_{ij} for gene j, respectively.

Alternative control methods To evaluate the performance of the autoencoder in a broader picture, we implemented two alternative state-of-the-art methods namely PEER[Stegle et al., 2012] and PCA.[Wold, 1966; Oba et al., 2003; Troyanskaya et al.,

2001] The two methods were used in replacement of the autoencoder to control the covaration in the data. In the case of PCA, we obtained the matrix of expected counts by setting the encoder and decoder matrices $\mathbf{W}_e$ and $\mathbf{W}_d$ to the first q loadings while using the gene mean as bias term $\mathbf{b}$. In the case of PEER, we set the number of factors to one-fourth of the number of samples as suggested by Stegle et al. [2012]. We then subtracted the residuals from the log-transformed counts and multiplied the size factors to obtain c_{ij}. For PEER, we used the provided residuals to compute Z scores to avoid numerical inaccuracies due to conversion to counts. For both PCA and PEER, we fitted a NB model with an additional per-gene adjustment parameter a_j, which captures deviations between the estimated mean from the log-normal and NB model, to obtain NB P values. Specifically:

$$\mathrm{P}(k_{ij}) = \mathrm{NB}\left(k_{ij} | \mu_{ij} = a_j \cdot c_{ij}, \theta_j\right). \tag{3.13}$$

3.4 Controlling gene expression for confounding effects

In the Kremer dataset, we already observed a strong correlation structure and showed that controlling manually for some covariats can lower the correlations (Fig. A.3). Also in GTEx, we observed correlation structures in each tissue (Fig. 3.2c and 3.3a). While some of the structures in the GTEx tissues can be explained by known covariates such as the sequencing center and Hardy scale classification, unknown confounders are still present in the data (Fig. 3.2c).[Li et al., 2017] Additional unknown sources of variations in RNA-seq data can arise from origin of the tissue, population structure, or hidden confounders such as poorly understood systematic technical variations. Applying the autoencoder on the counts allowed covariations to be estimated and controlled for across all GTEx tissues (Fig. 3.3a). The autoencoder even managed to remove almost all correlation structures across all GTEx tissues without prior knowledge by reducing the correlation from 0.20 ± 0.06 to 0.04 ± 0.01 (mean of the mean absolute sample-wise correlation across all GTEx tissues $\pm$ standard deviation). The dimension q of the autoencoder was fitted for each dataset with a denoising autoencoder by selecting the dimension maximizing the area under the precision-recall curve for identifying corrupted counts. This resulted in an estimated latent space dimensions q of 45 and 21 for the suprapubic skin tissue and the Kremer dataset, while using the PCA approach yielded 54 and 24, respectively. In general, q increased proportional with the sample size, while the number of expressed genes did not have an impact on the dimension (Fig. 3.3b-c). Changing the corruption amplitude and scheme had little impact on the optimal dimension. Only a very low amplitude required a higher dimension in the experimental datasets, probably due to the fact of genuine outliers present already in the data (Supplemental Fig. S3 in Brechtmann et al. [2018]).

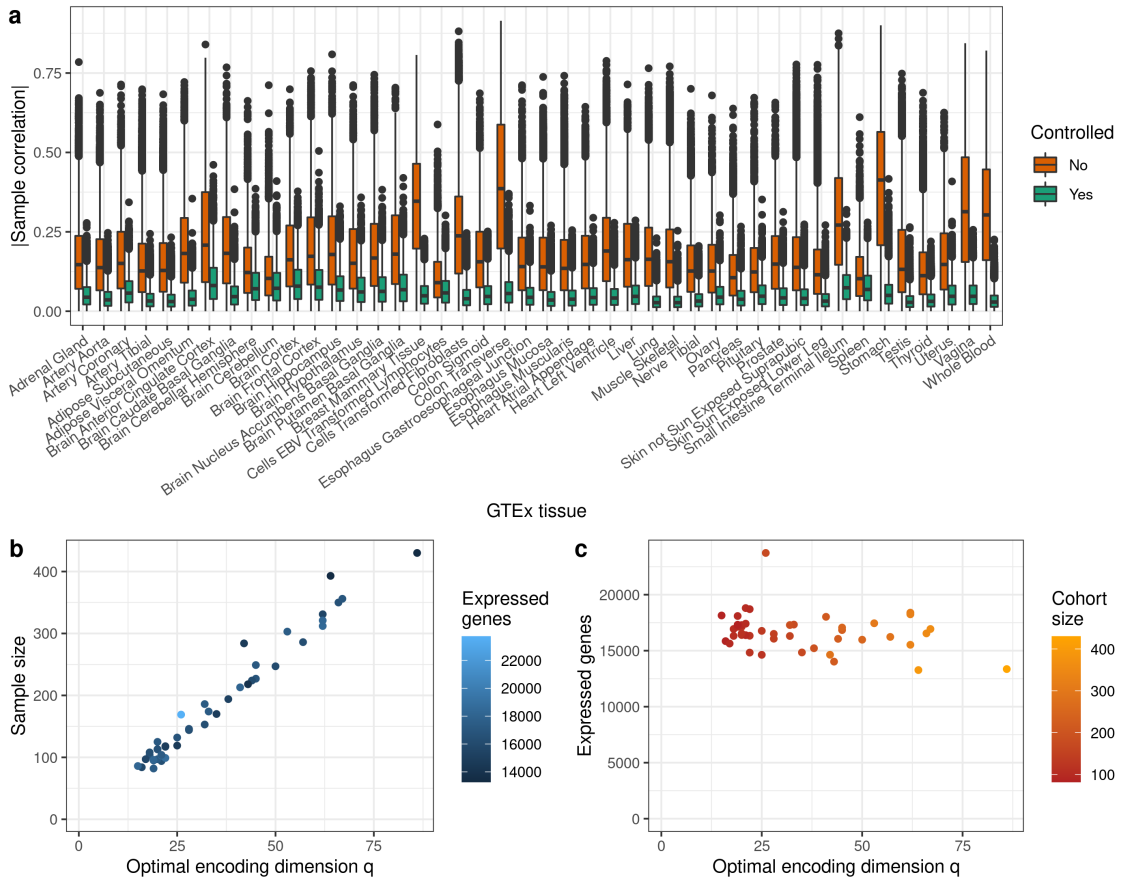

Figure 3.3: Controlling for known and unknown confounders in GTEx tissues. (a) Boxplots of absolute values of between-sample correlations of gene-centered log-transformed read counts for 48 GTEx tissues before (orange) and after (green) correction for the latent space. **(b)** Optimal encoding dimension q (x-axis) plotted against the cohort size (y-axis). The lighter the color the higher the number of expressed genes in the given tissue. Each point represents a GTEx tissue. **(c)** Same as b, but where the number of genes is on the y-axis and the color encodes the cohort size.

3.5 Detection of expression outliers with the negative binomial distribution

After managing to remove confounders in the count data with different approaches, we aimed to understand the impact of the modeling on the performance of outlier detection. Working with count data from RNA-seq, we assume that the counts follow a NB distribution where the mean is provided by the given model as a count expectation.[Whitaker, 1914; Love et al., 2014; Robinson et al., 2010]. Using the expectations c_{ij} and a gene-specific dispersion parameter θ_j, expression outliers are detected if the observed counts are significantly deviating from these expected values. While our autoencoder approach uses a NB loss function, PCA and PEER assume a normal distribution and therefore

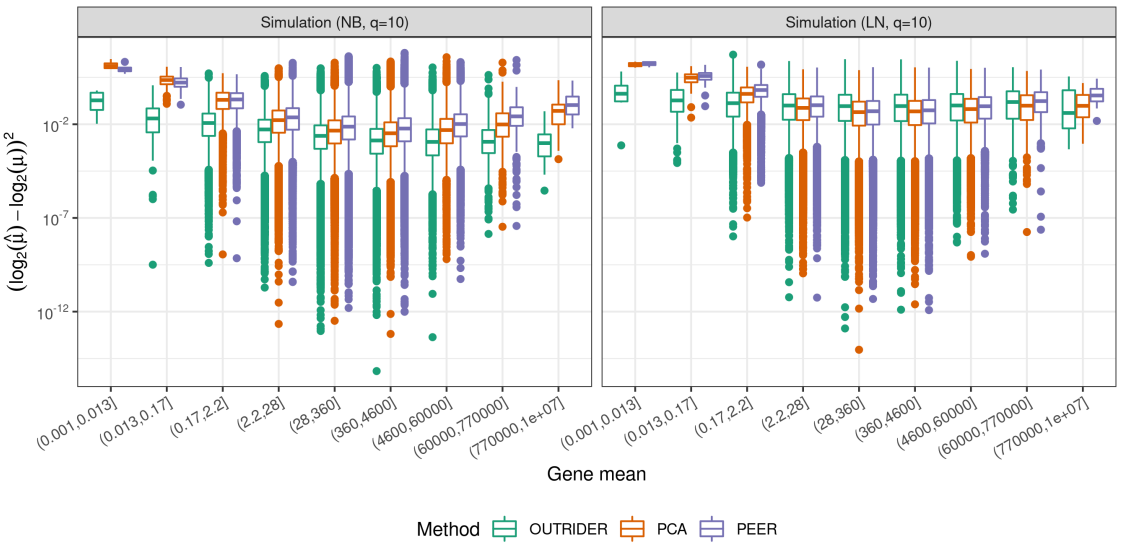

Figure 3.4: An NB-based autoencoder recovers better expected counts on simulated data than log-normal-based models. Boxplots of squared differences between expected counts and simulated means in $\log_2$ space binned into 9 logarithmically spaced mean gene expression bins for OUTRIDER, PCA, and PEER on simulated data. The data is stratified by the simulation scheme: negative binomial and log-normal, respectively. Adapted from Brechtmann et al. [2018].

require some transformation of the count data, typically the log-transformation. To this end, we used the two NB and log-normal based simulated datasets (Section 3.2) following the OUTRIDER and the PCA/PEER assumptions, respectively. The expected counts fitted by OUTRIDER were closer to the simulated means than the fitted expectations by PCA and PEER in the NB simulations across the count spectrum (Fig. 3.4. On the log-normal simulated data, the algorithms performed similar well. Notably, the autoencoder model outperformed PCA and PEER on the lower and higher end of simulated counts in both cases. These observations emphasize the relevance of using a count distribution for fitting the expected counts, especially in the low count range.

Investigating the quantile-quantile plots per gene indicated that our autoencoder modeled the data well even in the presence of outliers (Fig. A.7) across datasets rendering the P values usable for outlier detection. By exchanging the underlying model with PCA and PEER, the P values decreased across the datasets potentially inflating the type I error (Fig. 3.5A-B). In line with these findings, the detected expression outliers per sample at an FDR < 0.05 was more uniform for the autoencoder approach than for PCA and PEER for both datasets (Fig. 3.5C-D). In addition, the autoencoder produced no aberrant sample compared to PCA and PEER. Aberrant samples were defined as a sample having more than 0.5% expression outlier genes. Accumulated over all GTEx tissues, we found 9, 18, and 214 out of 8,166 samples to be aberrant by using the autoencoder, PCA, and PEER as model, respectively. While in most cases all three methods had similar results, the autoencoder did not find any outlier genes in some samples with

a high number of outlier genes called by PCA and PEER (Fig. 3.5E-F). Overall, this demonstrates that our OUTRIDER implementation of combining an autoencoder with a NB significance test is appropriate for detecting expression outlier genes in RNA-seq data. It also highlights the importance of working directly on count data with the appropriate NB distribution, rather than assuming a log-normal distribution that requires transformation of the input data.

3.6 Benchmarking gene expression outlier detection methods

While developing a method, one has to evaluate its performance and also compare it with the state-of-the-art alternatives. Unfortunately, in the case of rare diseases and outlier detection, there was neither a gold standard dataset nor a benchmark scheme that was available. Therefore, we developed a process to assess the sensitivity and specificity by injecting artificial outliers into experimental data while monitoring the performance of recalling them. We note that this approach is underestimating the performance, as experimental data can include already genuine outliers.[Kremer et al., 2017; Li et al., 2017; Ferraro et al., 2020] We have implemented a similar approach as for the denoising autoencoder (Section 3.3). Specifically, with a pre-specified amplitude on the logarithmic scale (z scores), we injected outlier counts k^o_{ij} with a frequency of 10^{-4} as:

$$k^o_{ij} = \text{round}\left(s_i \cdot 2^{\bar{u}_j \pm e^z \sigma_{u_j}}\right), \tag{3.14}$$

where $\bar{u}_j$ is the mean of u_{ij} (Eq. 3.8) for gene j in the log space.

We applied this benchmark strategy to the GTEx suprapubic skin tissue. To disentangle the impact of the amplitude and directionality of the aberrant events on the outlier detection performance, we injected outlier counts according to three scenarios with four different amplitudes ($z \in \{2, 3, 4, 6\}$): (i) only underexpression, (ii) only overexpression, and (iii) equally distributed under- and overexpression. This resulted in 381 injected outliers across all samples. We benchmarked OUTRIDER against seven different algorithms and monitored for each the recall of injected read-count outliers and the precision. The precision was defined as the number of injected outliers among the reported outliers for each method. The applied methods included OUTRIDER, PCA, and PEER ranked by NB based P values and z scores. For completeness, we included Cook's distance and Pearson residuals as implemented in DESeq2[Love et al., 2014] and edgeR,[Zhou et al., 2014] respectively, normalized with the available covariates of sex, age, and ischemia time. The precision-recall curves showed that the P value based OUTRIDER ranking outperformed ranking by z scores, except in the case of simulated outliers with a high amplitude ($z = 6$, Fig. 3.6). Notably, the two commonly used z score cutoffs $|z| > 2$[Li et al., 2017; Frésard et al., 2019] and $|z| > 3$[Cummings et al., 2017] recalled almost all the outliers (median = 97%) regardless of the method, but at the cost of a high FDR (precision < 0.05). The NB P value based methods performed similar with a slight advantage for OUTRIDER towards outliers with a smaller amplitude. Using a P value

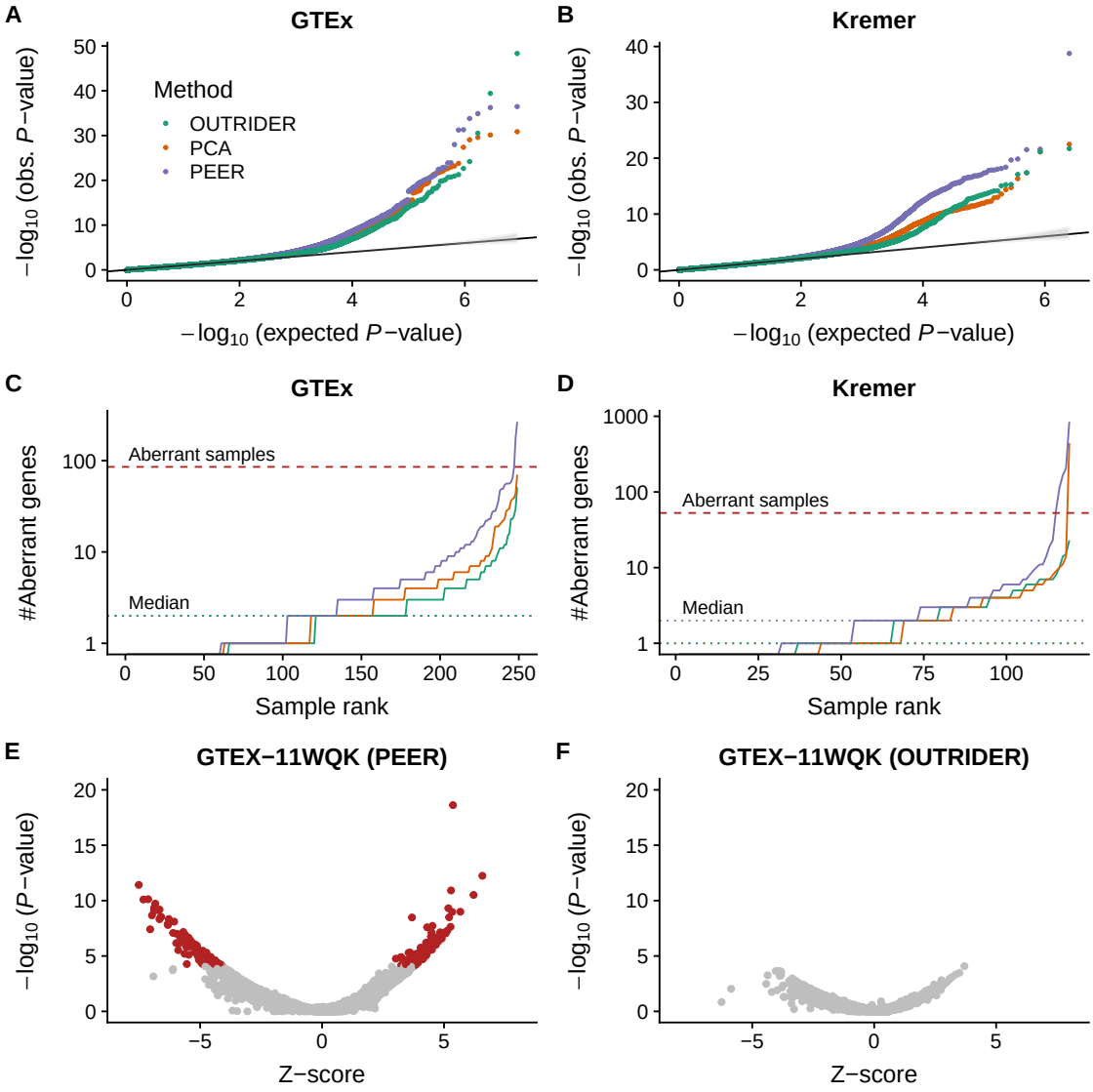

Figure 3.5: RNA-seq expression outlier detection (A) Quantile-quantile plot for the GTEx dataset. Observed P values are plotted against the expected P values for three different methods. The diagonal marks the expected distribution under the null hypothesis with 95% confidence bands (gray). **(B)** Same as A but for the Kremer datasets. **(C)** Number of aberrantly expressed genes (FDR < 0.05) per sample for the data shown in A. The dashed line represents the abnormal sample cutoff ($> 0.5\%$ aberrantly expressed). **(D)** Same as C but for the data in B. **(E)** P values versus z scores for a representative abnormal sample in PEER. Genes with significantly aberrant read counts are marked in red. **(F)** Same as E but controlled with the autoencoder in OUTRIDER. Adapted from Brechtmann et al. [2018].

based cutoff (FDR < 0.05) increased the precision up to 0.52 especially for OUTRIDER in the scenario of under expression outliers, but at the cost of reduced recall. Ranking

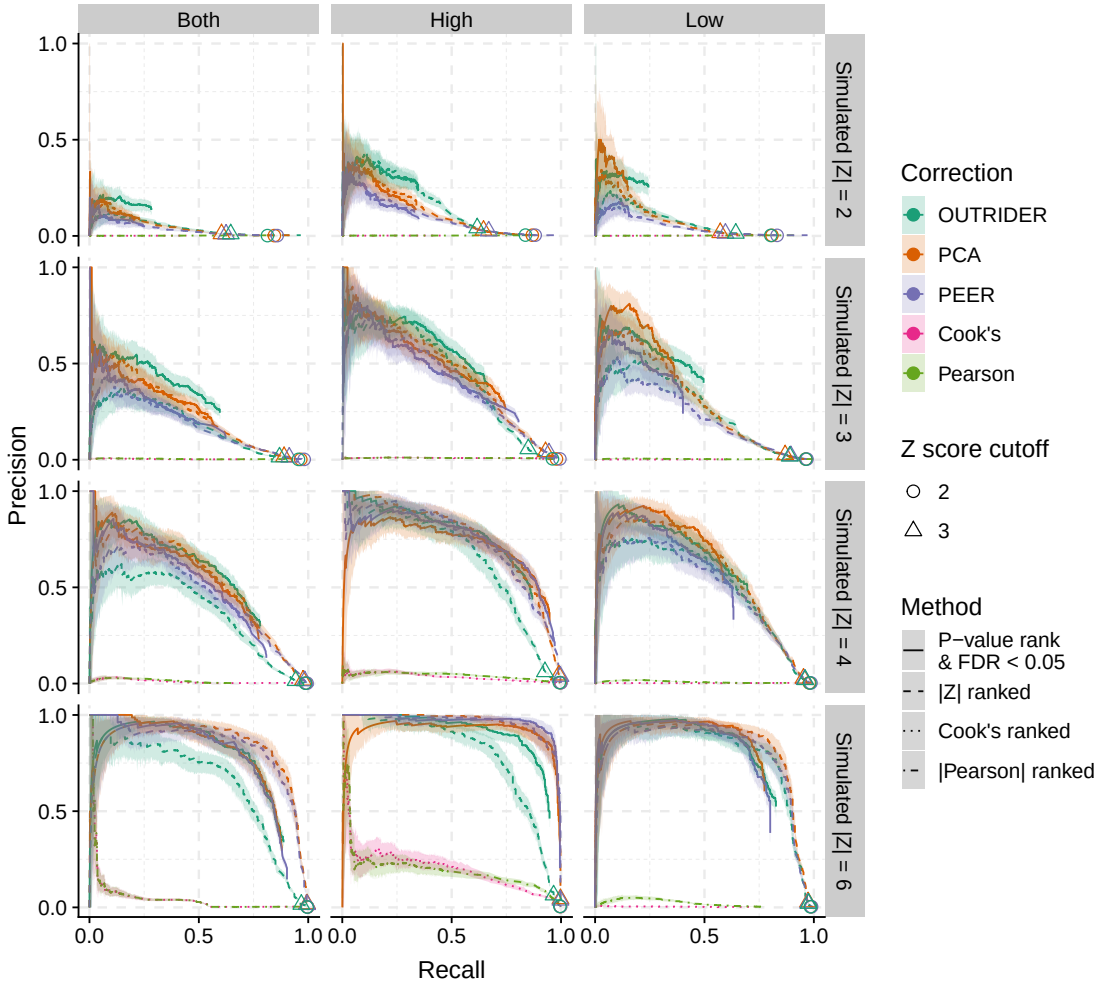

Figure 3.6: Outlier detection benchmark in GTEx. The proportion of simulated outliers among reported outliers (precision) plotted against the proportion of reported simulated outliers among all simulated outliers (recall) for 8 different ranking methods. The 8 ranking methods are OUTRIDER (green solid), PCA (orange solid), and PEER (blue solid) sorted by P value with FDR < 0.05, OUTRIDER (green dashed), PCA (orange dashed), and PEER (blue dashed) sorted by z score, DESeq2 normalization with known covariates sorted by Cook's distance (pink dotted), and DESeq2 normalization with known covariates sorted by absolute value of Pearson residuals (olive green dashed and dotted). Plots are provided for four simulated amplitudes (by row, with simulated absolute z scores of 2, 3, 4, and 6, top to bottom, respectively) and for three simulation scenarios (by column for aberrantly high and low counts, for aberrantly high counts only, and for aberrantly low counts only, left to right, respectively). The ranking of outliers was bootstrapped to obtain 95% confidence areas. Adapted from Brechtmann et al. [2018].

by Cooks' distance and Pearson residual performed poorly in all cases, rendering them

as inadequate alternatives for the detection of expression outliers in rare diseases. To understand why some events were not recovered by the P value based methods, we stratified the precision-recall curves by the mean expression levels of the genes. As expected, this revealed that outliers in genes with low expression levels are difficult to call. Still OUTRIDER recalled 39% of the events with a precision of 0.49 for genes with a mean expression below 58 outperforming the z score ranking (Fig. A.8), which is in line with our results from the simulations (Fig. 3.4). Applying the same benchmark strategy on the Kremer dataset yielded similar results while injecting 113 artificial outliers (Fig. S7 in Brechtmann at al).[Brechtmann et al., 2018] Altogether, this analysis delineates the importance of using a count distribution and a P value based strategy in two ways: (i) P values provide a principled way to establish a cutoff that accounts for statistical significance and multiple testing and (ii) z scores can be unstable, especially for genes with low expression levels.

3.7 Rare variant enrichment in GTEx

Another way of benchmarking a method is by evaluating the results in its domain specific context. To this end, we performed an enrichment of rare variants among outliers. The underlying assumption is that if you observe a non-synonymous rare variant in a gene, you are likely to observe aberrant expression in that gene than if you do not have such a rare variant. This assumption is inline with results by Zeng et al. [2015]; Li et al. [2017], where rare variants were linked to aberrant gene expression in humans. To obtain the set of potentially expression changing variants, we selected from the GTEx WGS data (V7)[The GTEx Consortium et al., 2015] only rare variants with a MAF < 0.05 within the 635 GTEx samples as well as in gnomAD.[Karczewski et al., 2020] In addition, we filtered for variants predicted to have moderate or high impact according to the Variant Effect Predictor (VEP).[McLaren et al., 2016] To make our analysis comparable with Li et al. [2017], we selected the same 441 individuals to compute the enrichment score. The enrichment was computed for rare variants found within outlier genes as the proportion of outliers having a rare variant over the proportion of non-outliers having a rare variant as described by Li et al. [2017].

On all GTEx tissues, we applied OUTRIDER, PCA, and PEER. We computed for the rare variant enrichment for three P value cutoffs and for three z score cutoffs. For all cutoffs, OUTRIDER achieved the highest enrichment compared to the alternative approaches, regardless of whether P values or z scores were used (Fig. 3.7). As expected, the enrichment correlated positively with the stringency of the cutoff. Interestingly, for the P value based enrichment OUTRIDER performed even better with a more stringent cutoff compared to PCA and PEER. Together with the benchmark results, this indicates that OUTRIDER can not only detect expression outliers, but that its results can be associated with genetic variants and thus interpreted biologically.

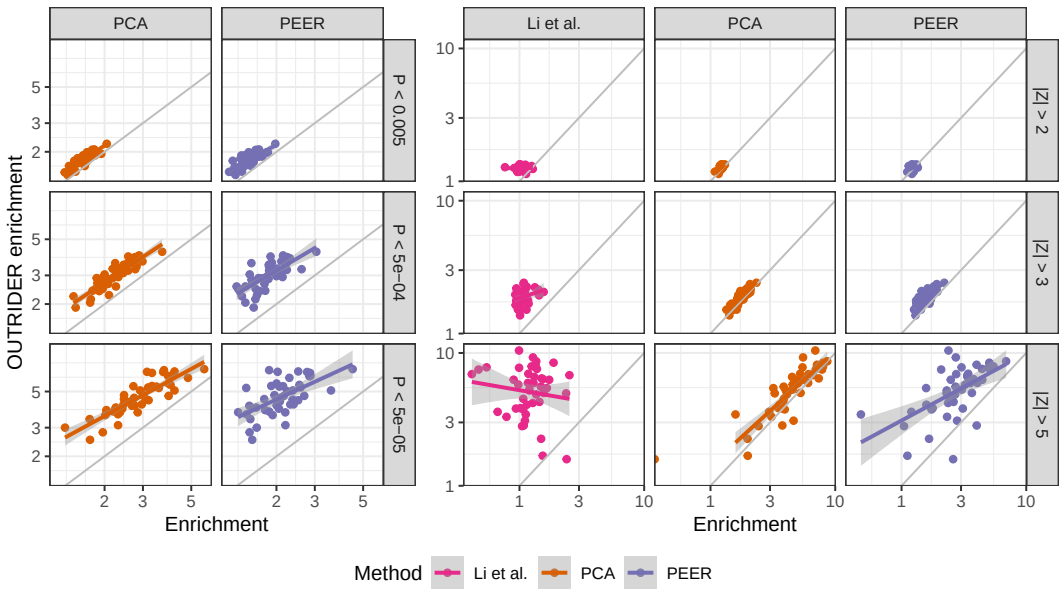

Figure 3.7: Expression outlier based rare variant enrichment in GTEx. Enrichment of rare (MAF < 0.05), moderate, and high impact variants (according to VEP[McLaren et al., 2016]) computed on genes found to be aberrantly expressed using OUTRIDER plotted against enrichments computed on genes found to be aberrantly expressed using z scores published by Li et al. [2017], PCA, and PEER for all GTEx tissues using three P value and z score cutoffs. Adapted from Brechtmann et al. [2018].

3.8 Reanalysis of the Kremer dataset using OUTRIDER

As the ultimate aim of OUTRIDER is to be applied in diagnostics of rare diseases, we used the Kremer dataset as a reference. Applying OUTRIDER resulted in a recall of 61 events (9.9%) identified by the adapted DESeq2 approach in the previously 48 undiagnosed individuals (Section 2.2.1). Although OUTRIDER detected 85 new expression outliers, none of them led to a new diagnosis. Interestingly, OUTRIDER was able to recall all six pathogenic events, even if they were not identified as such a priori (three expression outliers, one mono-allelic expression, and two splicing defects, Fig. 3.8). By identifying the misspliced *CLPP* and *MCOLN1* as expression outliers show the increased sensitivity of OUTRIDER and the importance of controlling for confounding effects. Replacing the autoencoder with PCA and PEER resulted in 3.8 and 7.8 times more outliers while missing 2 and 1 pathogenic events, respectively (Kremer et al. [2017] Fig. S9).

As in rare diseases often only a few samples are available, it is crucial to understand the impact of the sample size on the sensitivity. To this end, we randomly subsetted the Kremer dataset, retaining the six pathogenic events and monitoring their normalized P values. As expected, the P values were negatively correlated with the sample size (Fig. 3.8). While the autoencoder approach needed all 119 samples to recall all six events, 60

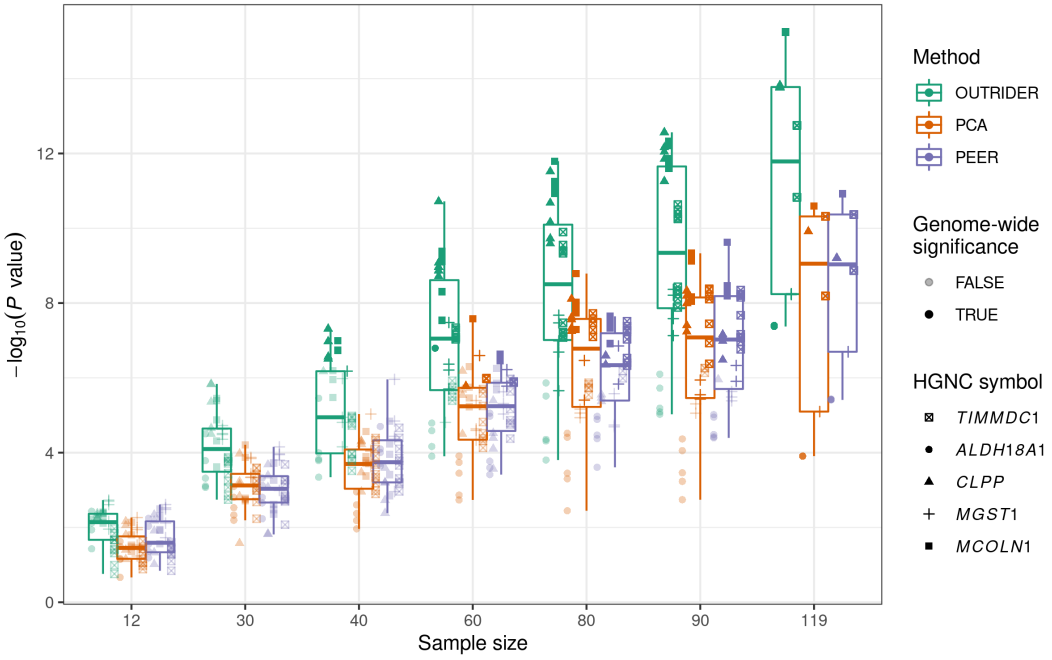

Figure 3.8: Sample size analysis. Negative $\log_{10}$ P values are plotted against the number of samples in the subset of the Kremer dataset, for the 6 pathogenic genes validated by Kremer et al. [2017]. For each subset size, five random sets of samples containing the samples with the known outliers were drawn. Genes that are genome-wide significant (FDR < 0.05) are marked darker. Adapted from Brechtmann et al. [2018].

samples were enough to recall 2/3 of the events missing the 50% reduced *ALDH18A1* and the second *TIMMDC1* case. PCA and PEER in contrast needed 90 samples to recall at least 2/3 of the known cases.

3.9 Short summary

We have introduced here OUTRIDER, a software package for detecting aberrant gene expression in RNA-seq data in the context of rare disease diagnostics. It uses a denoising autoencoder scheme to provide expected expression levels while automatically controlling for known and unknown confounders. OUTRIDER uses a NB distribution to test for significance based on expected counts and a gene-wise dispersion parameter. Further, we introduced a benchmark strategy to evaluate the performance of expression outlier detection by injecting artificial outlier counts into experimental data. In addition, we evaluated OUTRIDER's performance by computing enrichments for rare variants in GTEx. OUTRIDER outperformed alternative methods in both benchmarks by achieving better precision recall curves and by having higher enrichments. Overall, we highlight the importance of assessing the significance by using P values over z scores and demonstrate that OUTRIDER is capable to retrieve biologically relevant information. Through the

relevance of OUTRIDER for the diagonstic of rare diseases and the packaging of the software into an R/Bioconductor package, we foresee that it will be implemented to support RNA-seq based rare disease diagnostics.

4 Detection of aberrant splicing events in RNA-seq data with FRASER

The methodology, results, and figures presented in this chapter are part of the manuscript "Detection of aberrant splicing events in RNA-seq data with FRASER" from Mertes et al. [2021]. The author's contributions are included in it. In short, I conceived the method together with Julien Gagneur. I developed the software and analysed the data together with Ines Scheller. The loss functions and the corresponding gradients were mainly derived by Ines Scheller. The manuscript was written by me, Ines Scheller, and Julien Gagneur.

4.1 Motivation

Genetic variants affecting splicing are known to be a major cause of rare diseases. It is estimated that up to 30% of Mendelian disease causing variants are linked to or causing aberrant splicing.[López-Bigas et al., 2005; Wang and Cooper, 2007; Park et al., 2018b; Taylor and Sobczak, 2020; Ellingford et al., 2019] The underlying mechanisms include exon skipping, truncation, and elongation but also intron retention.[Wang and Burge, 2008; Scotti and Swanson, 2015] The variants that affect splicing can be located anywhere from donor and acceptor sites to known regulatory splicing elements to deeply intronic regions. With the increased implementation of WGS based diagnostics, deep intronic variants are now easily identifiable, but most of them are just ignored or still remain as VUS as their prioritization is still limited.[Jian et al., 2014; Jaganathan et al., 2019] Even advances in predicting the effect of variant effects on splicing with machine learning techniques did not overcome these challenges.[Xiong et al., 2015; Rosenberg et al., 2015; Cheng et al., 2019; Jaganathan et al., 2019; Cheng et al., 2021] Therefore, genetic diagnosis guidelines require additional functional evidence to classify a variant as pathogenic.[MacArthur et al., 2014; Richards et al., 2015] One such functional evidence can be provided by RNA-seq. Over the last years, RNA-seq has been proven effective in detecting disease-causing aberrant splicing events.[Kremer et al., 2017; Cummings et al., 2017; Frésard et al., 2019; Gonorazky et al., 2019; Murdock et al., 2021]. RNA-seq can not only be used to validate or invalidate a link of a VUS to aberrant splicing[Cummings et al., 2017] but also to call de novo aberrant splicing events transcriptome-wide, including the activation of deep intronic cryptic splice sites.[Kremer et al., 2017; Cummings et al., 2017; Gonorazky et al., 2019; Murdock et al., 2021]

At the time, three methods have been developed to diagnose rare disease by detecting aberrant splicing events in RNA-seq.[Kremer et al., 2017; Cummings et al., 2017; Frésard et al., 2019; Gonorazky et al., 2019] In addition, SPOT[Ferraro et al., 2020] and LeafCutterMD[Jenkinson et al., 2020] were developed recently as dedicated splicing outlier detection tools. All five methods rely on the split reads to identify splicing events. Split reads are RNA-seq reads aligned to a given chromosome with at least one gap that usually matches an annotated intron. Calling aberrant splicing beyond annotations is important as creations of novel splice sites have a strong pathogenic potential by leading to frameshifts, ablation of protein sequences, or creation of nonfunctional protein sequences. The method developed by Cummings et al. [2017] and adapted by Gonorazky et al. [2019] applied only cutoffs on absolute and relative split read counts which has the limitation of not assessing the statistical significance of the events nor being data driven. In contrast, we assess the significance of aberrant splicing by adapting LeafCutter[Li et al., 2018] to splicing outlier calling (Section 2.2.2). This approach, along with the more recent LeafCutterMD[Jenkinson et al., 2020] and SPOT,[Ferraro et al., 2020] which are also multivariate approaches, allowed controlling for FDR but not for possible covariation structure. Frésard et al. [2019] showed that strong covariations do exists in split-read-based splicing metrics and that it is beneficial to control for them by performing a PCA. The origins of such covariations can be the same as for gene expression like sex, genetic variation, RIN, and technical biases. Splicing outliers where then identified by a z score cutoff ($|z| \geq 2$) making it impossible to control for FDR. Despite the development of multiple approaches to detect aberrant splicing, each method had limitations. In addition, by focusing solely on split reads, these methods are all blind to intron retention as intron retention reduces the number of split reads covering the given intron, but not the splice ratios.

Therefore, we developed FRASER (Find RAre Splicing Events in RNA-seq) to address the limitations in aberrant splicing detection by using splice ratios and splicing efficiency as metric, providing an automated mechanism to control for known and unknown confounders in them, and assessing each event for significance based on the BB distribution (Fig. 4.1). We applied the same strategy as in OUTRIDER (Section 3.3) by using an autoencoder to control for covariations. Extensive benchmarking using the GTEx dataset shows the advantage of FRASER over the alternative methods. The clinical relevance of using FRASER for aberrant splicing detection is highlighted by detecting a splice defect in *TAZ* by reanalysing the Kremer dataset.

4.2 Statistical modeling

Learning from previous studies,[Kremer et al., 2017; Cummings et al., 2017; Frésard et al., 2019; Brechtmann et al., 2018] we aimed to tackle three major limitations, which are similar to the ones described in OUTRIDER (Section 3.3). First, as split reads alone can not detect intron retention, we used the intron-centric splicing efficiency (θ) in addition to the percent spliced in (ψ) metric.[Pervouchine et al., 2013], Second, we em-

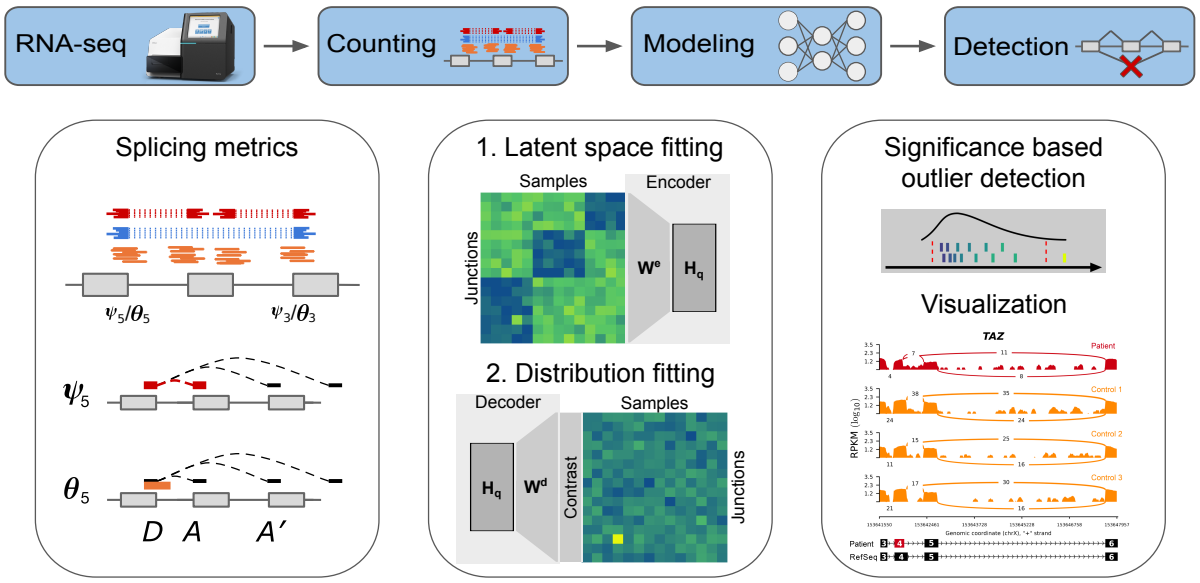

Figure 4.1: The FRASER aberrant splicing detection workflow. The workflow starts with RNA-seq aligned reads and performs splicing outlier detection in three steps. First (left column), a splice site map is generated in an annotation-free fashion based on RNA-seq split reads. Split reads supporting exon–exon junctions as well as non-split reads overlapping splice sites are counted. Splicing metrics that quantify alternative acceptors (ψ_5), alternative donors (ψ_3), and splicing efficiencies at donors (θ_5) and acceptors (θ_3) are then computed. Second (middle column), a statistical model is fitted for each splicing metric that controls for sample covariations and overdispersed count ratios. Third (right column), outliers are detected as data points that deviate significantly from the fitted model. Candidates are then visualized using a genome browser. D donor site, A acceptor site. Made in ©BioRender - `biorender.com`. Adapted from Mertes et al. [2021].

ployed a autoencoder scheme to control for covariation the splice metrics. And third, we assessed the statistical significance of aberrant splicing events using the BB distribution.

Read counting and splicing metrics. Introns and their respective acceptor and donor splice sites were defined by the data instead of relying on existing annotations. Using an annotation-free approach enables the detection of novel introns. Introns were defined by the alignment gaps of the RNA-seq split reads. Split reads were extracted from the BAM files and counted using the R/Bioconductor packages GenomicAlignments and GenomicRanges.[Lawrence et al., 2013]. For non-strand specific RNA-seq data the strand of the given intron was predicted by the dinucleotides of the splice site.[Reyes et al., 1996]. Based on the set of acceptor and donor site, non-split reads overlapping the splice sites were counted to compute the splicing efficiency to enable the detection of intron retention. Specifically, the non-split reads were counted for each splice site using the R/Bioconductor Rsubread package.[Liao et al., 2019]. For non-split reads, we requiring an alignment of at least 5 nt on each side of the splice site for robustness

against mapping errors of very short overhangs, as described by Braunschweig et al. [2014]. Using the split reads, we computed the intron-centric percent spliced in metrices ψ_5 and ψ_3 as described by Pervouchine et al. [2013] per sample according to Equation 1.1 and 1.2, respectively. To enable the detection partial or full intron retention, we used the splice site specific splicing efficiency metric θ_5 and θ_3 computed the same way as ψ_5 and ψ_3 according to Equation 1.4 and 1.5, respectively.[Pervouchine et al., 2013] While we calculated θ specifically for the donor and acceptor site, θ_5 and θ_3 were not distinguished later in the modeling step. Therefore, we call the splicing efficiency metrics jointly as θ. To remove noise from the read count data and to improve the modeling, we applied two filters: (i) we kept only introns supported by at least 20 split reads in at least one sample and (ii) we removed introns from the analysis where more than 95% of the samples had zero coverage.

Autoencoder-based beta-binomial hypothesis testing The metrics ψ_5, ψ_3, and θ are count proportions. For each of these metrics, we model the distribution of the numerator conditioned on the value of the denominator using the BB distribution. As we used the NB distribution for gene read counts to account for overdispersion (Section 3.3), we use the BB distribution instead of the binomial distribution. As we model each metric the same way, we use the term ψ to refer to ψ_5, ψ_3, and θ. Specifically, for a given ψ metric, we assume that the split read count k_{ij} of the intron $j = 1, \ldots, p$ in sample $i = 1, \ldots, N$ follows a BB distribution with an intron-specific intra-class correlation parameter ρ_j and a sample- and intron-specific proportion expectation μ_{ij}:

$$\mathrm{P}(k_{ij}) = \mathrm{BB}\left(k_{ij} | n_{ij}, \mu_{ij}, \rho_j\right), \tag{4.1}$$

where n_{ij} defines the total number of split reads having the same donor site than intron j. μ_{ij} and ρ_j are both limited to the range $[0, 1]$. A more detailed parametrization of here used BB distribution can be found in the Supplementary Note 3 of Mertes et al. [2021].

The proportion expectations μ_{ij} are modeled with an autoencoder of encoding dimension $1 < q < min(p, N)$ that captures covarations beween samples similar to the autoencoder in OUTRIDER (Eq. 3.2-3.7). Specifically, we model:

$$\mu_{ij} = \sigma(y_{ij}) = \frac{\exp(y_{ij})}{1 + \exp(y_{ij})}, \tag{4.2}$$

$$\mathbf{y}_i = \mathbf{h}_i \mathbf{W}_d + \mathbf{b}, \tag{4.3}$$

$$\mathbf{h}_i = \widetilde{\mathbf{x}}_i \mathbf{W}_e, \tag{4.4}$$

where the vectors $\mathbf{h}_i$ are the rows of the matrix $\mathbf{H}$, the $N \times q$ projection of the data onto the q-dimensional latent space, $\mathbf{W}_e$ is the $p \times q$ encoding matrix, $\mathbf{W}_d$ is the $q \times p$ decoding matrix, and the p-vector $\mathbf{b}$ is a bias term. The input vector $\widetilde{\mathbf{x}}_i$ is given by the

intron centered logit-transformed pseudocount ratios. It is defined as:

$$\widetilde{x}_{ij} = x_{ij} - \bar{x}_j, \tag{4.5}$$

$$\bar{x}_j = \text{mean}_i(x_{ij}), \tag{4.6}$$

$$x_{ij} = \text{logit}\left(\frac{k_{ij}+1}{n_{ij}+2}\right), \tag{4.7}$$

$$\text{logit}(a) = \log\left(\frac{a}{1-a}\right). \tag{4.8}$$

We add a pseudocount to prevent the division by zero in Eq. 4.7 and hence add 1 to the nominator while adding 2 to the denominator. The combination of equations 4.2-4.8 is what we call the autoencoder implementation in FRASER, in short, $\mu_{ij} = \text{AE}_{\text{FRASER}}(k_{ij})$.

Fitting the autoencoder and beta-binomial distribution parameters In our statistical model, we have four parameters to fit: (i) $\mathbf{W}_e$ and (ii) $\mathbf{W}_d$, the en- and decoding matrix, (iii) $\mathbf{b}$ the bias term, and (iv) $\boldsymbol{\rho}$ the intra-class correlations per gene of the BB distribution. The fitting of these parameters is achieved in two steps. In the first step, $\mathbf{W}_e$ and $\mathbf{W}_d$ is fitted using a PCA on $\widetilde{\mathbf{X}}$ as implemented in the pcaMethods package.[Stacklies et al., 2007] Specifically, we set $\mathbf{W}_e$ and the transposed $\mathbf{W}_d$ to be the first q loadings of the PCA and the bias term $\mathbf{b}$ is set to $\bar{\mathbf{x}}$. In the second step, the intra-class correlation parameters ρ_j are fitted per intron j given the count proportion expectations μ_{ij} using a BB loss function. Specifically, we use the *optimize* function from R[Team, 2021] and minimize the average negative BB log-likelihood in parallel across introns. Detailed derivations of the used loss functions and the respective gradients can be found in the Supplemental Notes 3 of Mertes et al. [2021].

Finding the optimal encoding dimension To find the optimal hyperparameter q, the latent space dimension, we applied again a denoising autoencoder scheme as done in OUTRIDER (Section 3.3). As the fitting takes more time for the BB distribution and here only the hyperparameter q is optimized, we subsetted first the input matrix $\widetilde{\mathbf{X}}$ to 15,000 random introns out of the top 30,000 most variable introns having at least a mean total coverage of 5. Then we injected with a frequency of 10^{-2} corrupted read counts k^o_{ij} into the data and measured the performance of the autoencoder to recall them. We injected the corrupted read count as described in the next paragraph 4.2. Finally, the optimal dimension q is then selected as the dimension maximizing the area under the precision-recall curve for identifying the corrupted read counts.

In silico injection of artificial outliers. To fit the FRASER hyperparameter as well as to compare the splicing outlier detection performance between FRASER and other methods, we developed a procedure to inject artificial outliers into a given dataset. As the splice metrics derived from introns sharing the same donor or acceptor are dependent, we injected only one outlier per splice site and sample and we considered only positions

with a total coverage of at least 10 to be able to inject actual outliers. We injected corrupted read counts k^o_{ij} with a frequency of 10^{-2} and 10^{-3} for the optimization of the encoding dimension and the benchmark, respectively. The corrupted read count k^o_{ij} is injected by changing the original read count k_{ij} such that the value of ψ_{ij} changes by $\Delta\psi^o_{ij}$. We draw $\Delta\psi^o_{ij}$ from a uniform distribution as:

$$\Delta\psi^o_{ij} \sim \pm \mathrm{U}(0.2, \Delta\psi^{\max}_{ij}), \tag{4.9}$$

$$\Delta\psi^{\max}_{ij} = \begin{cases} \psi_{ij}, & \text{if down-regulation.} \\ 1-\psi_{ij}, & \text{if up-regulation.} \end{cases} \tag{4.10}$$

where $\Delta\psi^{\max}_{ij}$ is the maximal possible $\Delta\psi_{ij}$ for intron j in sample i dependent on the injection direction, which is random. To ensure that an aberrant splice ratio can be injected the direction is switched if $\Delta\psi^{\max}_{ij} < 0.2$. Taking the pseudocounts into account, the outlier count k^o_{ij} is then given by

$$k^o_{ij} = \mathrm{round}((\psi_{ij} \pm \Delta\psi^o_{ij}) \cdot (n_{ij}+2) - 1). \tag{4.11}$$

In order to provide a biologically realistic outlier injection scheme that preserves the total amount of reads, the counts for the introns l sharing the same donor or acceptor, respectively, with k^o_{ij} are changed accordingly, where the $\Delta\psi^o_{ij}$ change is distributed equally over all secondary introns l, as follows:

$$\Delta\psi^s_{il} = -\Delta\psi^o_{ij} \cdot \frac{\psi_{il}}{1-\psi_{il}} \quad \textit{and} \tag{4.12}$$

$$k^s_{il} = \mathrm{round}((\psi_{il} \pm \Delta\psi^s_{il}) \cdot (n_{il}+2) - 1). \tag{4.13}$$

P value computation For every pair of intron j and sample i, we test the null hypothesis that the count k_{ij} with n_{ij} trials follows a BB distribution as described by Equation 4.1. To detect down and up regulation of a given intron, we compute two-sided P values using the probability estimates μ_{ij} and the fitted intra-class correlation ρ_j as:

$$P_{ij} = 2 \cdot \min\left\{ \frac{1}{2}, \sum_{k=0}^{k_{ij}} \mathrm{BB}(k_{ij}|n_{ij},\mu_{ij},\rho_j), 1 - \sum_{k=0}^{k_{ij}-1} \mathrm{BB}(k_{ij}|n_{ij},\mu_{ij},\rho_j) \right\}. \tag{4.14}$$

Due to the nature of the discrete BB distribution, both one-sided P values can exceed $1/2$ at the same time, for which we introduced the $1/2$ term.

As already mentioned, introns sharing the same donor or acceptor are not independent. Therefor, we correct the P values for each splice site with the FWER using Holm's method, which holds under arbitrary dependence assumptions,[Holm, 1979] and report the minimal corrected P value per splice site. If gene-level P values are requested, an additional FWER step is performed at the gene level. On top of the FWER correction, we correct P values per sample for multiple testing using the FDR method transcriptome-wide. We use the Benjamini-Yekutieli procedure because it applies under

positive dependence,[Benjamini and Yekutieli, 2001] which is necessary because introns within and beyond a gene can be co-regulated and hence be correlated, even after controlling for covariations by the autoencoder.

Z score and $\Delta\psi$ calculation Z scores z_{ij} are computed per intron on the difference on the logit scale between the measured ψ_{ij} value including pseudocounts and the proportion expectation μ_{ij} as follows:

$$z_{ij} = \frac{\delta_{ij} - \text{mean}_i(\delta_j)}{\text{sd}_i(\delta_j)}, \tag{4.15}$$

$$\delta_{ij} = \text{logit}\left(\frac{k_{ij}+1}{n_{ij}+2}\right) - \text{logit}(\mu_{ij}), \tag{4.16}$$

The $\Delta\psi$ values are calculated as the difference between the observed ψ_{ij} value on the natural scale including pseudocounts and the proportion expectations μ_{ij}:

$$\Delta\psi_{ij} = \psi_{ij} - \mu_{ij} = \frac{k_{ij}+1}{n_{ij}+2} - \mu_{ij}. \tag{4.17}$$

Alternative control methods In addition to the above described autoencoder approach, we implemented two alternative approaches to fit the decoder matrix $\mathbf{W}_d$ and the bias term $\mathbf{b}$ given the latent space $\mathbf{H}$. We used an iterative approach similar to the one used by OUTRIDER (Section 3.3) to fit $\mathbf{W}_d$ together with $\mathbf{b}$ and $\boldsymbol{\rho}$ using a negative BB log-likelihood loss function. The initialization of the parameters is done as described above. As the parameters of $\mathbf{W}_d$, $\mathbf{b}$, and $\boldsymbol{\rho}$ are independent across introns, they can be optimized in parallel. We start by optimizing ρ_j given the decoder coefficients w_j^d and the bias b_j (step 1). Subsequently, we optimize w_j^d and b_j given ρ_j in step 2. Steps 1 and 2 are repeated until the average negative log-likelihood of each step in one iteration does not differ by more than the convergence threshold of 10^{-5} from the last step of the previous iteration, or until 15 iterations are reached. For the optimization, we use the L-BFGS method implemented in the R function *optim* to fit the decoder coefficients and the bias.[Byrd et al., 1995]

Since outlier data points can have a strong impact on the fit of a given distribution, downweighting such outlier data points during the fit can make it more robust. In the diagnostic setting, we do expect to have such events and therefore, we incorporated weights as described by Zhou et al. [2014] into the negative BB log-likelihood loss function. Specifically, we defined the weight w_{ij}^r for each observation based on its Pearson residual. The Pearson residual r_{ij} of the observed data point x_{ij} (Eq. 4.7) with respect to the BB distribution including the pseudocounts is defined as follows:

$$r_{ij} = \frac{\text{observerd} - \text{expected}}{\sqrt{\text{Var}(\text{expected})}} = \frac{x_{ij} - \mu_{ij}}{\sqrt{\frac{\mu_{ij}(1-\mu_{ij})(1+(n_{ij}-1)\rho_j)}{n_{ij}+2}}} \tag{4.18}$$

The weights w_{ij}^r for sample i and intron j are obtained from these residuals using the Huber function:[Huber, 1964]

$$w_{ij}^r = \begin{cases} 1, & \text{if} |r_{ij}| \geq k. \\ \frac{k}{|r_{ij}|}, & \text{otherwise.} \end{cases} \tag{4.19}$$

where we use $k = 1.345$ as suggested in the edgeR package,[Zhou et al., 2014] which leads to the downweighting of about 5% of the data points. These weights are then included in the calculation of the negative log-likelihood yielding the average weighted negative log-likelihood L^W:

$$L^W = \frac{1}{p \times N} \sum_{i,j} w_{ij}^r L_{ij}, \tag{4.20}$$

$$L_{ij} = -\log\left(\text{BB}(k_{ij} | n_{ij}, \mu_{ij}, \rho_j)\right), \tag{4.21}$$

where L_{ij} is the negative BB log-likelihood of sample i and intron j. Detailed derivations of the used loss functions and the respective gradients can be found in the Supplementary Note 3 by Mertes et al. [2021].

Alternative splicing outlier detection methods To evaluate the performance of the autoencoder in a broader picture, we implemented five alternative methods, namely (i) a naïve BB regression, (ii) a PCA-based z score approach similar to Frésard et al. [2019], (iii) the LeafCutter adaptation (Section 2.2.2), (iv) LeafcutterMD,[Jenkinson et al., 2020] and (v) SPOT.[Ferraro et al., 2020] The naïve BB regression served as baseline. The parameters were estimated with the VGAM package[Yee, 2015] in R and the data was not corrected for any covariates. PCA-based z scores were computed according to Eq. 4.15. Instead of regressing out the top q principal components accounting for 95% of the variation within the data as done by Frésard et al. [2019], only the top q principal components maximizing the precision-recall of in silico injected splicing outliers were used in the regression. In addition to the Leafcutter adaptation developed in Section 2.2.2, we applied the Dirichlet multinomial based approaches LeafcutterMD[Jenkinson et al., 2020] and SPOT[Ferraro et al., 2020] on the RNA-seq data as recommended with default parameters.

4.3 Controlling the splice metric with denoising autoencoders

To assess the performance of FRASER, we used the same GTEx dataset as used for the development of OUTRIDER (Section 3.2) resulting in 7,842 RNA-seq samples from 48 tissues of 543 assumed healthy donors after quality filtering. Although we assume that all donors are healthy, we expect the presence of aberrant splicing, just as we have observed aberrant gene expression (Section 3.5).[Li et al., 2017; Ferraro et al.,

2020] We quantified for all samples per tissue the intron-centric splice metrics. We chose the intron-centric splicing metrics ψ_5 and ψ_3[Pervouchine et al., 2013] over the exon-centric percent spliced in metric developed by Katz et al. [2015] as they can be quantified without prior knowledge of gene annotations and are capable of detecting novel introns. In addition, we used the splice efficiency metrics θ_3 and θ_3 to identify partial or complete intron retention events, as these values are lower when splicing is compromised. After filtering for expressed introns per tissue, we detected on average 137,058 ($\pm 5,848$ standard deviation across tissues) donor sites and 136,743 ($\pm 5,920$) acceptor sites (Fig. A.9). Less then 2% of the detected splice sites were not annotated by GENCODE (release 28).[Harrow et al., 2012] As with gene expression, we observed sample correlations in the intron-centered logit-transformed splice metrics for all GTEx tissues (Fig. 4.2a-c). Overall, the average absolute sample correlation was not as strong as with gene expression with 0.10 instead of 0.20 (± 0.05 and ± 0.06 standard deviation across tissues, respectively, Fig. 4.2d). Sample correlations were tissue specific and were detected across all splice metrics. Increasing the mean expression filter on the introns for the hierarchical clustering increased the correlations ruling out an impact of pseudocount on the correlations. While some of the clusters matched with existing annotations like the RIN (heart) or death classification (blood), not all of the clusters could be explained by known covariates. These is consistent with Frésard et al. [2019] observations and highlights the importance of controlling for confounders that can arise from common genetic variation, sex, technical artifacts, or other unknown factors.

To control for those between-sample covariations, we applied our autoencoder approach per tissue and splice metric by modeling a low-dimensional latent space in the logit space. As we optimize the encoding dimension q by maximizing the area under the precision-recall curve of recalling injected aberrant splicing counts, we investigated the impact of different injections scenarios. Using a fixed $\Delta\psi$ for injection showed that a higher dimension is needed for smaller amplitudes (Fig. A.10a). We followed the approach of OUTRIDER (Section 3.3) and used a uniform distribution for the $\Delta\psi$ values to be independent of the amplitude. The choice of q was robust as each of the injection scenarios plateaued around the optimal encoding dimension. Using the uniform injection approach resulted in an optimal encoding dimension of 15, 16, and 12 for ψ_5, ψ_3, and θ, respectively, on average across the GTEx tissues. As for gene expression, the optimal encoding dimension correlated positively with the number of samples within the tissue (Fig. A.10b). Controlling for the latent space, we showed that the autoencoder managed to remove almost all correlation structures across all GTEx tissues without prior knowledge by reducing the absolute between-sample correlations from 0.10 ± 0.05 down to 0.02 ± 0.01 (mean $\pm$ standard deviation across tissues, Fig. 4.2d).

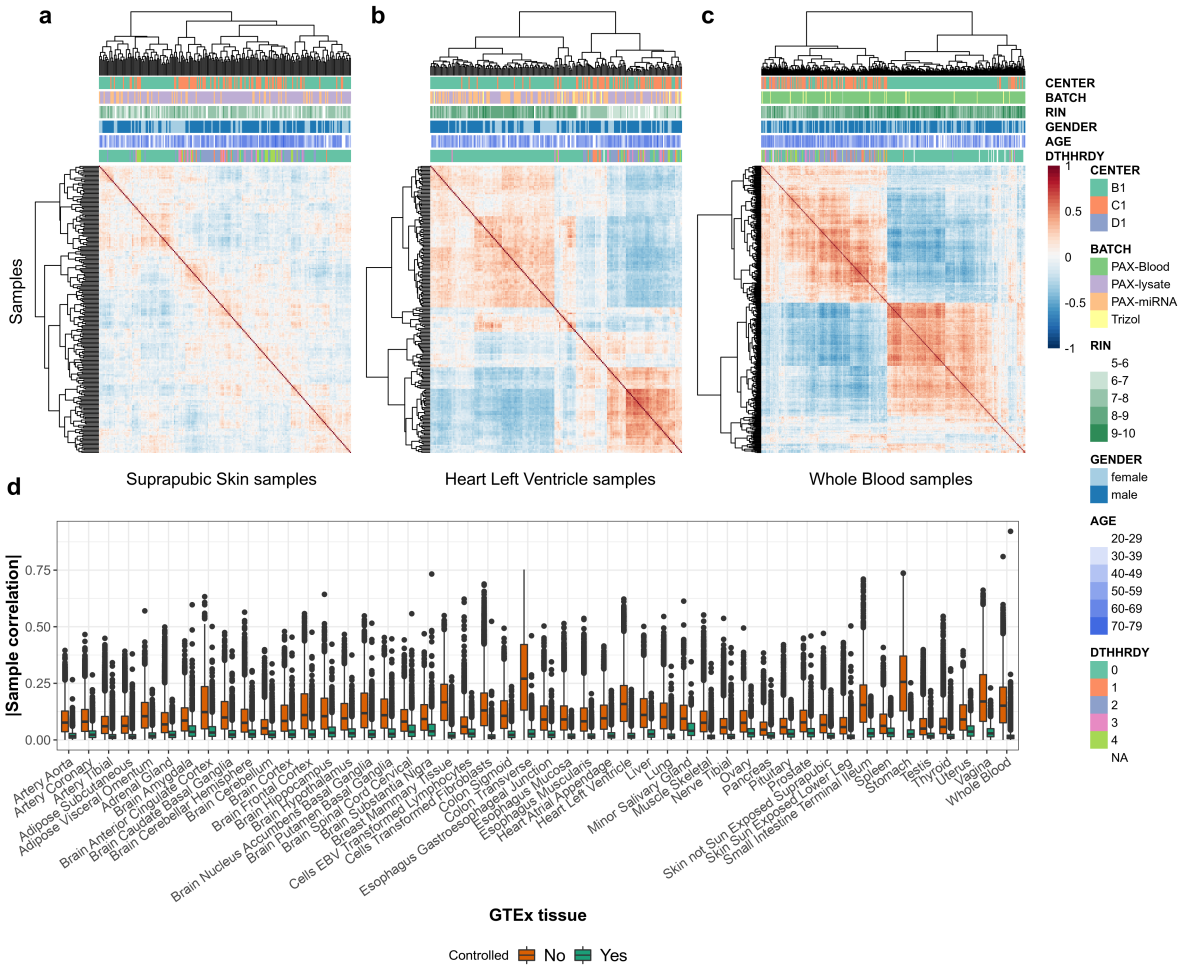

Figure 4.2: Tissue-specific correlation structure for ψ_3 (a) Intron-centered and logit-transformed ψ_3 values of the 10,000 most variable introns clustered by samples (columnsand rows) for the GTEx suprapubic skin tissue (n=222). Red and blue depict relative high and low intron usage, respectively. Colored horizontal tracks display sequencing center, batch, RNA integrity number (RIN), gender, age, and cause of death (DTHHRDY, Hardyscale classification) of the samples. **(b)** Same as a but for the left ventricle heart tissue (n=211). **(c)** Same as a but for the whole blood tissue (n=369). **(d)** Boxplots of absolute values of between-sample correlations of row-centered logit-transformed ψ_3 for 48 GTEx tissues before (orange) and after (green) correction for the latent space. The intron-centered ψ_3 values were clipped to the [0.01,0.99] interval before logit-transformation. Adapted from Mertes et al. [2021].

4.4 Detection of aberrant splicing events with the beta-binomial distribution

After establishing the effectiveness of the autoencoder in removing between-sample covariations, we investigated the impact of the model on the detection of aberrant splicing events using a BB distribution to assess significance. Assessing the significance in splicing data can be done in two ways: (i) using a Dirichlet-multinomial distribution per gene and (ii) using a BB distribution per intron. The rational of using a Dirichlet-multinomial distribution per gene is that all introns within a gene are dependent as ratios of each splice sites are defined by the expression ratios of the different isoforms. The drawback of this is that the introns have to be assigned to genes by annotation or clustered if an annotation-free approach is used as done in LeafCutter, LeafcutterMD, and SPOT. [Li et al., 2018; Jenkinson et al., 2020; Ferraro et al., 2020] To reduce the complexity, we chose to test each intron assuming a BB distribution using the expected count ratios modeled by the autoencoder. To detect up and down regulation of introns, we computed two-sided P values for each observation using the Equation 4.14. To have only one P value per splice site, we corrected them with Holm's FWER method.[Holm, 1979] To demonstrate the applicability of the BB distribution, we used again the GTEx suprapubic skin tissue as an example as done in the expression outlier analysis. Investigating the quantile-quantile plots per intron indicated that our autoencoder modeled the data well even in the presence of outliers (Fig. 4.3a-d) rendering the P values suitable for splicing outlier detection. Despite the spread in the usage of the 17th intron of *SRRT* across the population, which is captured by μ_{ij} and ρ_j, the aberrant splicing count is detected as significant ($P = 5.83 \times 10{-}11$, Fig. 4.3c-d). Looking at all introns per splice metric, P values tended to be conservative in general (Fig. 4.3e). Since we expect genuine aberrant splicing in the data,[Ferraro et al., 2020] it is expected that we do see an excessively low P value in every ten thousandth test. Running the same analysis across the GTEx tissues, resulted in similar observations.

4.5 Benchmarking aberrant splicing detection methods by in silico injection

To assess the performance of FRASER in general and in contrast to other methods, we followed the same strategy developed for the expression outlier benchmark (Section 3.6. We used the GTEx suprapubic skin tissue to inject aberrant splicing counts with a frequency of 10^{-3}, uniformly drawn amplitude, and random direction yielding 25,988, 26,153, and 49,169 outliers for ψ_5, ψ_3, and θ, respectively. Since we jointly model θ_5 and θ_3, we have twice as many events for θ compared to ψ_5 and ψ_3. Using this injection schema, we benchmarked FRASER against four different approaches and hence, monitored for each the recall of injected aberrant splicing counts and the precision. The precision was defined as the number of injected outliers among the reported outliers for each method. The applied methods included FRASER, a naïve BB regression, a

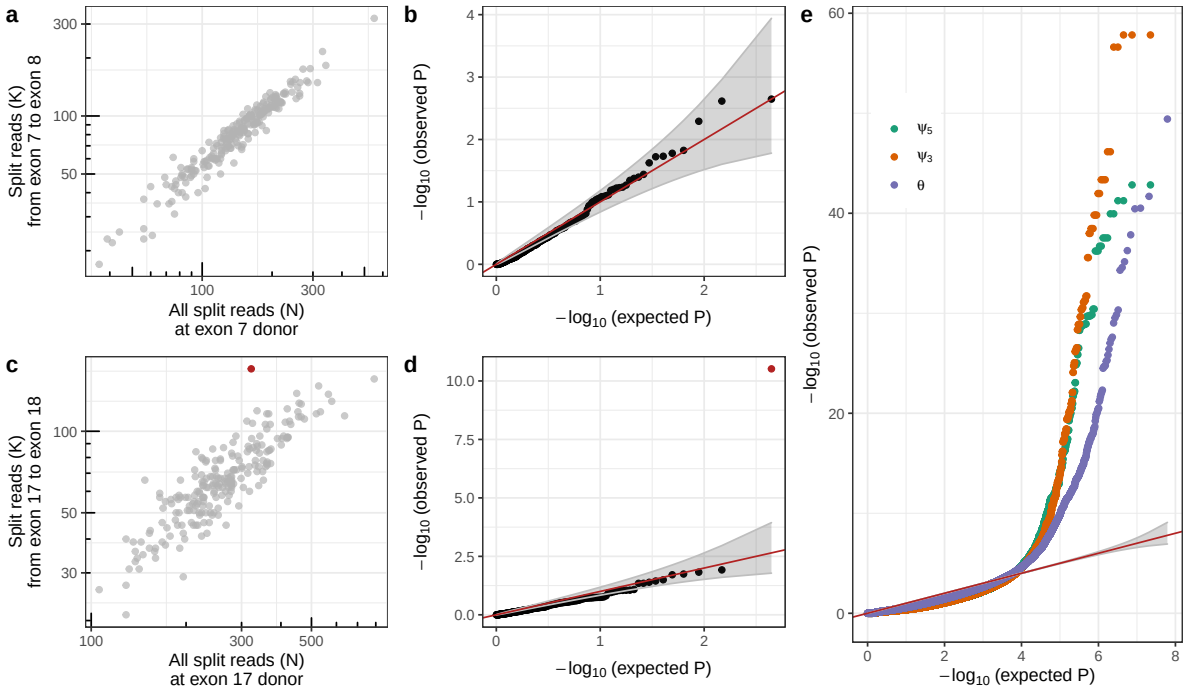

Figure 4.3: Splicing outlier detection based on the beta-binomial distribution. (a) Intron split read counts (y-axis) against the total donor split read coverage (x-axis) for the seventh intron of *SRGAP2*. **(b)** Observed negative log-transformed P values (y-axis) against expected ones (x-axis) of the ψ_5 metric for the data shown in a. Under the null hypothesis, the data are expected to lie along the diagonal (red, 95% confidence bands in gray). **(c)** Same as a for the 17$^{\text{th}}$ intron of *SRRT*, showing an outlier (FDR < 0.1, red). **(d)** Same as b for the 17$^{\text{th}}$ intron of *SRRT*. The outlier is marked in red. **(e)** Same as b across all introns and splice sites for ψ_5 (green), ψ_3 (orange), and splicing efficiency (θ, purple). a-e are based on the suprapubic skin tissue from GTEx (n=222). Adapted from Mertes et al. [2021].

PCA with BB regression using P values and z scores, and a pure PCA using z scores. The benchmark results showed three interesting points (Fig. 4.4). First, the naïve BB approach, which does not account for any covariation structure in the data, performed purely across all benchmark scenarios. Second, methods using P values to assess significance while controlling for covariation outperformed z score based methods. And third, using a z score cutoff equal to 2 over a FDR based cutoff equal to 0.1 yielded two orders of magnitude more outliers and a drastic drop in precision (5% vs. 92% with FRASER) for a small increase in recall (98% vs. 84% with FRASER (Fig. 4.4b). These results are resembling the findings from the gene expression outlier detection benchmark and demonstrate the strong advantage of assessing the significance using an appropriate count distribution over absolute z scores cutoffs.

We used this benchmark not only to evaluate the performance of FRASER but to investigate alternative ways of fitting the decoder given the latent space. Specifically,

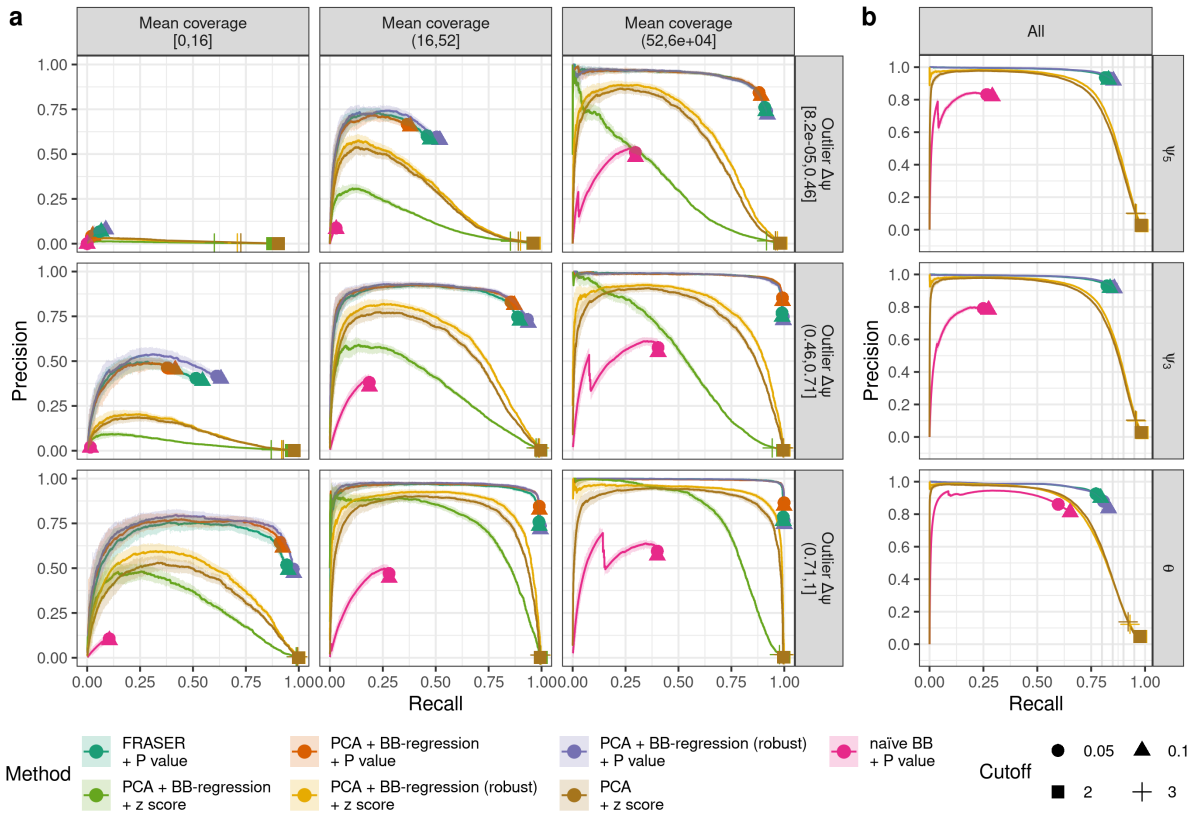

Figure 4.4: Splicing outlier detection benchmark in the GTEx suprapubic skin tissue. (a) The proportion of simulated outliers among reported outliers (precision, y-axis) plotted against the proportion of reported simulated outliers among all simulated outliers (recall, x-axis) for different aberrant splicing detection methods (color) for the ψ_5 metric only. All events with $|\Delta\psi| < 0.1$ are ranked last. Plots are stratified equally by injected amplitudes ($\Delta\psi$, by row) and junction coverage (by column). The points indicate commonly applied cutoffs (FDR < 0.1 and < 0.05 and absolute z scores > 2 and > 3). The darker lines mark the precision-recall curves computed for the full dataset while the light ribbons around the curves depict 95% confidence bands estimated by bootstrapping. **(b)** Same as a but stratified by splice metrics and not binned. Adapted from Mertes et al. [2021].

we investigated 3 methods, namely a least squares regression and a naïve and robust BB regression. Interestingly, all methods achieved similar performance when combining it with a significance test with the naïve BB regression having just a slight lower recall (Fig. 4.4a). Combining the different controlling methods with a z score approach showed a drastic drop of precision for the naïve BB regression as it was too sensitive to outliers in the data. Despite the small advantage of the robust BB regression over the least squares regression of logit-transformed splicing metrics, we opted for the latter to estimate the expected count ratios μ_{ij} due to the computational speed of the least squares regression.

FRASER tests each splice site independently. In principle, looking at all introns within a gene simultaneously could provide greater statistical power. This can be achieved by modeling the data using a Dirichlet-multinomial distribution as done in LeafCutterMD,[Jenkinson et al., 2020] SPOT,[Ferraro et al., 2020] and our LeafCutter adaptation (Section 2.2.2). To fairly assess this, we benchmarked the methods by injecting aberrant splicing events by swapping out all read counts of a given sample and gene from the brain cortex tissue into the suprapubic skin tissue. First, we selected 40 random individuals sequenced in both tissues. On those samples, we applied LeafCutter[Li et al., 2018] with the default parameters to detect alternative splicing. From the top 100 LeafCutter hits, we selected 60 random sample-gene pairs and injected for each pair the read counts from the brain into the skin tissue. We applied then all methods on the altered suprapubic skin tissue. Despite the theoretical advantage of the Dirichlet-multinomial approaches, FRASER outperformed them, as shown by the precision-recall curves (Fig. A.11). This could be due to the fact that no method, except FRASER, is offering to control the data for confounders.

4.6 Rare splicing variant enrichment in GTEx

Another way to benchmark a method is to evaluate the results in their domain-specific context. We used the same strategy as for the gene expression analysis in Section 3.7. To this end, we performed an enrichment of rare variants that potentially affect splicing among splice outliers. The underlying assumption is that if one observes a rare variant in the conserved splicing region or a rare variant predicted to affect the splicing of a gene, one is more likely to observe aberrant splicing in that given gene than if no such rare variant is observed. This assumption is inline with results by Ferraro et al. [2020], where rare variants were linked to aberrant splicing in humans. We created two sets of potentially splicing affecting variants. First, we selected from the GTEx WGS data (V7)[The GTEx Consortium et al., 2015] only rare variants with a MAF < 0.05 within the 635 GTEx samples as well as in gnomAD.[Karczewski et al., 2020] For the first set, we used VEP[McLaren et al., 2016] to filter for variants located around a splice site, specifically 1-3 bases within the exon and 1-8 bases within the intron. For the second set, we applied MMSplice[Cheng et al., 2019] on all rare variants ± 100 bp around an annotated splice site and filtered for those predicted to affect splicing ($|\Delta\text{logit}(\psi)| > 2$). We found on average 299.4 ± 207.6 and 66.0 ± 48.0 (mean $\pm$ standard deviation) variants for the splice site and the MMSplice variant set. We then computed enrichments as in Section 3.7. As splice affecting variants can have long ranging effects within a gene,[Drexler et al., 2020] we computed the enrichment on the gene level. Specifically, we computed enrichments for rare splicing variants found within splicing outlier genes as the proportion of outliers having a rare splicing variant over the proportion of non-outliers having a rare splicing variant.

To compare FRASER on the gene-level with the Dirichlet-multinomial based methods, we computed gene-level P values for FRASER using an extra FWER correction across all splice sites within a gene. For all four methods we computed enrichments

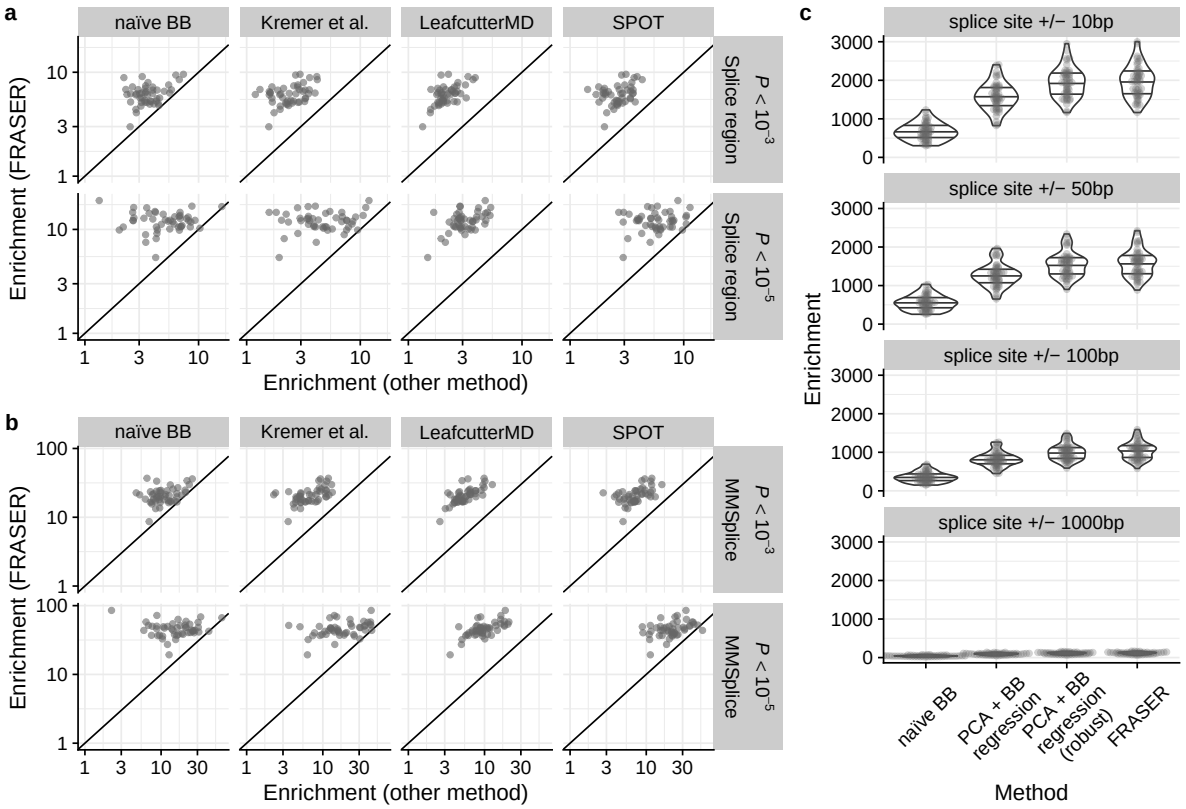

Figure 4.5: Enrichment for rare variants predicted to affect splicing. (a) Enrichment using FRASER (y-axis) against enrichment (x-axis) using different aberrant splicing detection methods (columns) for rare variants located in a splice region. The enrichment is calculated for different nominal P value cutoffs (rows). The applied methods are a naïve beta-binomial regression, the LeafCutter adaptation (Section 2.2.2), LeafCutterMD,[Jenkinson et al., 2020] and SPOT.[Ferraro et al., 2020] Each dot represents a GTEx tissue ($n = 48$). **(b)** Same as a but the enrichment is computed for rare variants predicted to affect splicing by MMSplice.[Cheng et al., 2019] **(c)** Violin plots of splice-site-based rare MMSplice variant enrichments (x-axis) for different correction methods (y-axis) and various variant range cutoffs (facets). BB beta-binomial. Adapted from Mertes et al. [2021].

across the 48 GTEx tissues. By achieving higher enrichments across the GTEx tissues, variant sets, and different nominal P value cutoffs, FRASER outperformed all methods including our LeafCutter adaptation, LeafCutterMD, SPOT, and non-corrected BB P values (Fig. 4.5a-b). Notably, the MMSplice variant set showed 2–10 times higher enrichments across all methods compared to the splice region variant set. Using a splice site specific enrichtment based on the MMSplice variant set, achieved even higher enrichments (Fig. 4.5c). Overall, this emphasizes on the biological site the importance of considering exonic or deep intronic variants as potential splice-affecting candidates and on the statistical site the importance of controlling for covariations within the data.

4.7 Reproducibility of aberrant splicing events in GTEx

Assessing the reproducibility of aberrant splicing calls using existing datasets is hard, as most of the datasets do not contain replicates. As GTEx contains many samples across tissues from the same donor, one can use this as a proxy to estimate the reproducibility of detecting aberrant splicing. To this end, we selected from the GTEx dataset donors with samples being sequenced in at least 20 tissues resulting in 195 donors. To make sure that an aberrant splicing event can be detected we filtered for sample-gene pairs where the gene passed the filtering criteria in at least 10 tissues. We then classified a given aberrant splicing event reproducible if it reached a nominal P value of $p < 10^{-3}$ in one or more additional tissues. We then computed the rare splicing variant enrichment on the entire dataset, using only the reproducible aberrant splice events.

Interestingly, the vast majority of the aberrant splicing events were not reproducible in any other tissue in GTEx regardless of the method used (Fig. 4.6a-b). This observation was also reported by Ferraro et al. [2020] when applying SPOT on the same data tuned specifically for this purpose. Compared to the other methods, FRASER achieved the highest percentage of reproducible aberrant splicing calls with increased reproducibility for events with lower P values (Fig. 4.6b). For instance, FRASER had a reproducibility rate of 22% compared to SPOT with 11% forfor outlier calls at a nominal P value of 10^{-7} in a given tissue that was reproducible in at least one additional tissue with $p < 10^{-3}$. In addition, the enrichment analysis of rare variants potentially affect splicing revealed that the enrichment increased with increasing reproducibility. This was true for all methods (Fig. 4.6c). These results suggest that non-reproducible splicing outliers exhibit a higher false positive rate compared to reproducible splicing calls. However, manual investigation of such tissue-specific aberrant splicing events with the Integrative Genomics Viewer (IGV)[Robinson et al., 2011] confirmed the outlier calls. Therefore, to understand and estimate this excess of tissue-specific outlier calls, biological replicates are needed.

4.8 Reanalysis of the Kremer dataset using FRASER

We started the development of FRASER with the aim to boost RNA-seq based rare disease diagnostics by robustly identifying aberrant splicing events. To demonstrate the performance of FRASER in this context, we reanalyzed the Kremer dataset using FRASER. In the clinical context, results are often manually examined by filtering for phenotypically relevant genes. Hence, we suggest to work with gene-level rather than splice-site-level statistics because they are easier to handle and only later, when the actual splice defect has to be localized after the identification of a putative disease-causing gene, splice-site statistics are helpful. Another common filter is the effect size, as larger effects are more likely to have strong downstream effects potentially resulting in physiological changes. FRASER identified a median of 12, 7, and 10 genes with at least one aberrant splicing event per sample for ψ_5, ψ_3, and θ, respectively, at a significance level of FDR < 0.1 and an effect size $|\Delta\psi| > 0.3$ (absolute difference between

observed and expected value, Fig. 4.7a). These numbers are consistent with the results obtained for all 48 GTEx tissues using the same cutoffs. With a total of 1,666 events, FRASER reported slightly fewer aberrant splicing events compared with the LeafCutter adaptation (1,725 events, Section 2.2.2). Nevertheless, FRASER identified all novel pathogenic splicing events, including those detected by other means in our pilot study (Fig. 4.7b). In particular, the intron retention event in *MCOLN1* detected through MAE (Section 2.2.3) was missed by the LeafCutter adaptation because it does not consider non-split reads. Moreover, FRASER reprioritized an exon truncation event in *TAZ* which ultimately led to a genetic diagnosis that is discussed in the next Section 5.1. Overall, testing simultaneously for alternative splicing using the ψ_5 and ψ_3 metrics and splicing efficiency using θ increased the number of detected events on the gene-level by two-fold over testing for alternative splicing alone.

We further investigated the impact of sample size on sensitivity with the same strategy as for OUTRIDER (Section 3.8). The purpose of this analysis is to provide guidance on study design and required sample size in the clinical setting. To this end, we randomly subsetted the Kremer data while keeping the samples with the 13 known pathogenic splicing events. After applying FRASER to the subsets, we monitored the recovery rate of those 13 events and their nominal P value. As expected, we observed the same drop of recovery rate and higher P values with smaller sample sizes as in OUTRIDER (Fig. A.12). Only 30 samples were needed to recover on average 85% of the events (11 out of 13). While FRASER recovered all events already with 50 samples for some subsets, 100 samples were needed to robustly recover them every time.

Altogether, these findings show the importance of using both alternative splicing and splicing efficiency metrics as well as dedicated statistical models in RNA-seq based diagnostics to boost sensitivity. In addition, these results demonstrate that 30 samples can be enough to detect the majority of disease-causing splicing defects.

4.9 Short summary

We have introduced here FRASER, an R/Bioconductor package for detecting aberrant splicing events in RNA-seq data in the context of rare disease diagnostics. It builds upon the developed architecture in OUTRIDER. It uses a denoising autoencoder scheme to provide expected read count proportions while automatically controlling for known and unknown confounders. FRASER detects aberrant splicing using splice site metrics and a BB distribution with a splice-site specific dispersion parameter. Further, we adapted the benchmark strategy developed for gene expression outlier detection to aberrant splicing to benchmark FRASER and proved its superior over other approaches. Through the application of FRASER on the Kremer dataset, we not only established a new diagnosis but also highlighted the importance of integrating the splicing efficiency metric to detect intron retention events.

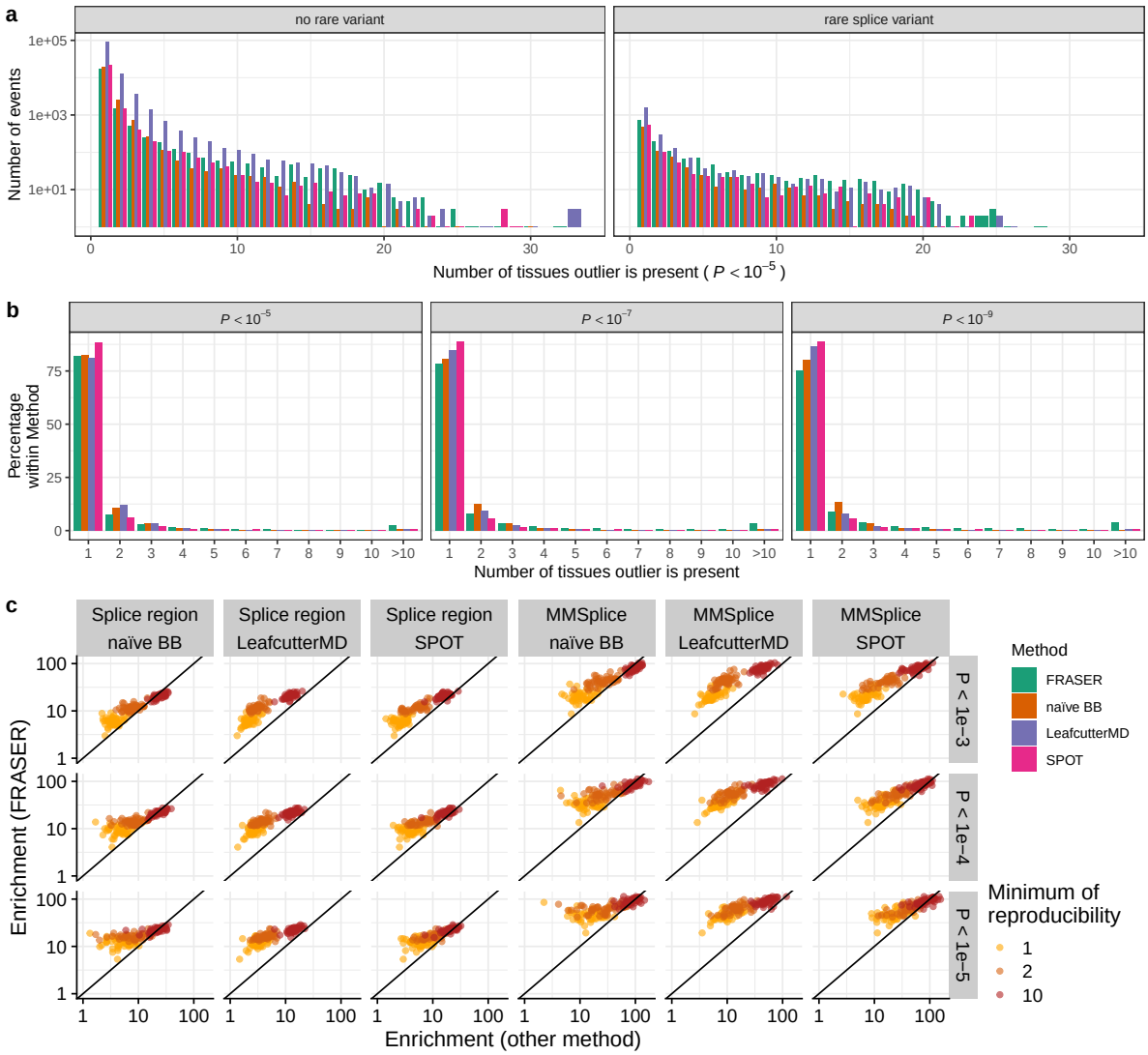

Figure 4.6: Reproducibility of splicing outlier calls and their rare splicing variant enrichment across GTEx tissues. (a) Barplot of the number of gene-level events (y-axis) against their reproducibility (x-axis) across GTEx tissues. The reproducibility is defined as the number of tissues an event is observed at a nominal P value $p < 10^{-3}$ given it was observed at least once at $p < 10^{-5}$. The data is stratified by associated variant status and grouped by the different methods: FRASER (green), naïve BB (orange), LeafcutterMD (purple), and SPOT (pink). **(b)** Same as a but plotted as the proportion (y-axis) of reproducible gene-level splicing outlier calls in GTEx tissues (number of tissues, x-axis) stratified by the primary outlier call cutoff. **(c)** Enrichment using FRASER (y-axis) against enrichment using the same methods as in a (x-axis, columns) stratified by the variant set (columns), namely rare splice site and MMSplice, respectively. The enrichment is calculated for different nominal P value cutoffs (rows) and increased reproducibility cutoff (color). Each dot represents a GTEx tissue. Adapted from Mertes et al. [2021].

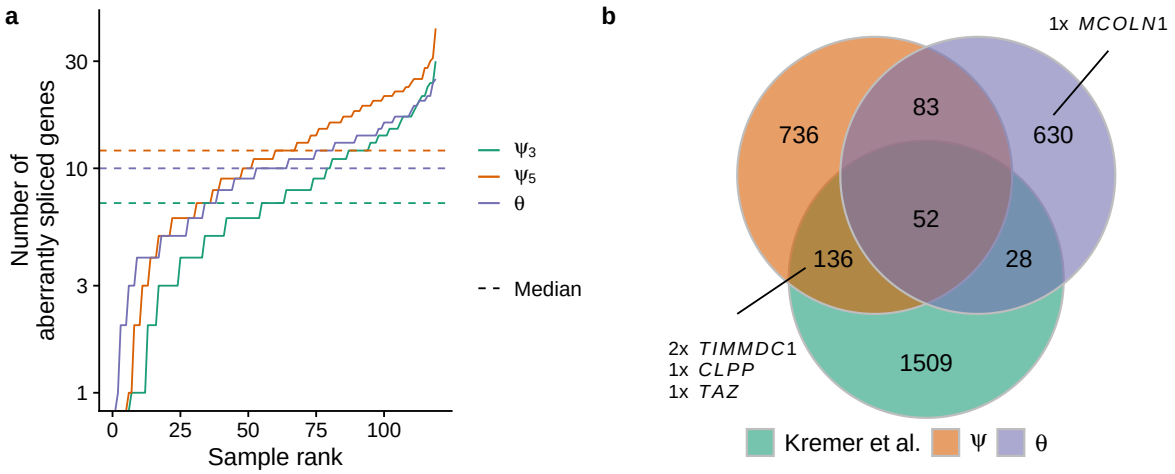

Figure 4.7: Aberrant splicing detection in the Kremer dataset. (a) Number of aberrantly spliced genes within the Kremer dataset (FDR < 0.1 and $|\Delta\psi| > 3$) per sample ranked by the number of events for ψ_5 (orange), ψ_3 (green), and θ (purple). **(b)** Venn diagram of the aberrant splicing events detected by FRASER using alternative splicing (orange, ψ) or splicing efficiency (violet, θ) only and detected by Kremer et al. [2017] (green). Pathogenic splicing events are labeled with the gene name. Adapted from Mertes et al. [2021].

5 Solving rare disease cases via RNA sequencing

Knowing is not enough; we must apply.
Willing is not enough; we must do.

Johann Wolfgang von Goethe

Content of this chapter is based on ongoing studies done jointly with Mirjana Gusic, Vicente Yépez, Robert Kopajtich with the guidance of Julien Gagneur and Holger Prokisch. The data presented is mostly unpublished. Clinicians from different institutes helped to gather the data. Mirjana Gusic, Robert Kopajtich, and Agnieszka Nadel prepared and sequenced the samples. Vicente Yépez helped to obtain and analyze the data. The results were interpreted jointly.

Since the initial pilot study in 2016, our rare disease cohort grew from 105 cases to currently 309 cases and is still growing. The methods developed in this thesis were continuously applied to every new incoming WES inconclusive case in a more and more standardized and automated way.[Yépez et al., 2021a] Ultimately, this RNA-seq based diagnostic strategy led to more than 40 molecular diagnoses in total over the last years.[Kremer et al., 2017; Mertes et al., 2021; Kopajtich et al., 2021; Yépez et al., 2021a] In this section, I will showcase three more cases in which RNA-seq was crucial to pinpoint the disease-causing variant or to prioritize candidate genes and to functionally interpret the underlying pathomechanism.

5.1 Exon truncation in TAZ caused by a synonymous variant

The reanalysis of the Kremer dataset with FRASER reprioritized a splicing defect in *TAZ*, Tafazzin. Specifically, it revealed a truncation of the fourth exon by aberrant alternative donor usage in individual #74116 ($\Delta\psi_3 = -0.88$ and FDR $= 1.98 \times 10^{-9}$, Fig. 5.1). Tafazzin catalyzes the maturation of cardiolipin, a major lipid constituent of the inner mitochondrial membrane that is involved in energy production and mitochondrial shape maintenance.[Houtkooper et al., 2009] A synonymous variant (c.348C>T) located 22 bp inside the fourth exon created a new upstream donor site. This homozygous variant in individual #74116 led to a nearly loss of wildtype *TAZ* by truncating the protein by

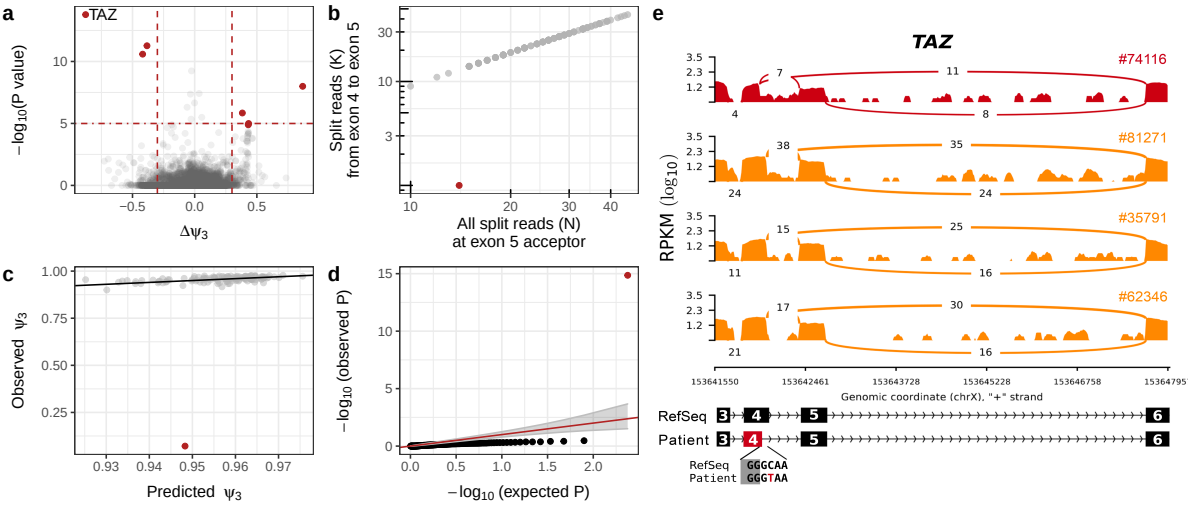

Figure 5.1: Detection of a pathogenic splicing defect in *TAZ* using FRASER. (a) Gene-level significance (-$\log_{10}(P)$, y-axis) versus effect ($\Delta\psi_3$, x-axis) for alternative donor usage for individual #74116. Six genes (red dots) passed both the genome-wide significance cutoff (horizontal dotted line) and the effect size cutoff (vertical dotted lines). **(b)** Number of split reads spanning from the fourth to fifth exon (y-axis) against the total number of split reads at the acceptor site of the fifth exon (x-axis) of *TAZ*. Sample #74116 (red) deviates from the cohort trend (red dot). **(c)** Observed (y-axis) against FRASER-predicted (x-axis) ψ_3 values for the data shown in b. **(d)** Quantile-quantile plot of observed P values (-$\log_{10}(P)$, y-axis) against expected P values (-$\log_{10}(P)$, x-axis) and 95% confidence band (gray) for the data shown in b. **(e)** Sashimi plot of the exon-truncation event in RNA-seq samples of the TAZ-affected (red) and three representative TAZ-unaffected (orange) individuals. The RNA-seq read coverage is given as the $\log_{10}$ RPKM-value (y-axis) and the number of split reads spanning an intron is indicated on the exon-connecting line. Underneath, the gene model of the RefSeq annotation is depicted in black and the aberrantly spliced exon is colored in red. The insert depicts the donor site-creating variant of the affected individual #74116. Adapted from Mertes et al. [2021].

8 amino acids (Fig. 5.1e). This event was overlooked in our pilot study even though our LeafCutter adaptation found it significant. It was not prioritized as it was not indexed in ClinVar[Landrum et al., 2018] at the time and due to its classification as synonymous. The reprioritization triggered a new literature survey, which revealed that Ferri et al. [2016] associated the same splicing defect in *TAZ* with cardiomyopathy at the same time as our pilot study. Cardiomyopathy is consistent with the myopathic facies and arrhythmias presented by individual #74116, thereby establishing the genetic diagnosis. Without the detection of the splicing defect in the RNA-seq data, the genetic variant would not have been reprioritized and clinically reclassified from VUS to pathogenic.

5.2 Identification of the expression of a pathogenic cryptic exon in MRPS30

In the second case, a newborn was presented with a mitochondrial complex V deficiency, cardiomyopathy, and metabolic acidosis. The young boy died in the same year due to heart failure. As WES on a skin biopsy sample did not reveal any candidates, RNA-seq was performed. OUTRIDER and FRASER both prioritized *MRPS30* as down-regulated gene expression and splicing outlier (Fig. 5.2). The overall expression of *MRPS30* was reduced by 75%. Inspecting the aberrant splicing calls, revealed the expression of a cryptic exon in the first intron. Variant calling on RNA-seq identified a homozygous deep intronic variant (c.602-468T>G) inside the cryptic exon that created a new acceptor side (Fig. 5.2e). Both SpliceAI (0.4 acceptor gain) and CADD (6.6) did not prioritized this variant as damaging. As *MRPL30* (OMIM #611838) is one of the 70 components of the mitochondrial ribosomes[Kenmochi et al., 2001] together with the matching phenotype and low percentage of wildtype expression, these findings are establishing the genetic diagnosis of individual #127272. As in silico prediction tools misclassified the variant by predicting no or low effect on translation while WES completely missed the variant, RNA-seq was needed to detect the expression of the cryptic exon and to interpret the functional effect of the deep intronic variant.

5.3 Mono allelic expression of a pathogenic missense variant in RRM2B

In the third case, a young boy was presented with general hypotonia, profound global developmental delay and sensorineural hearing impairment, periventricular white matter hyperdensities, and diffuse optic disc pallor. Biochemical analysis of the urine, revealed high lactic acid and protein levels and the excretion of glucose. WES prioritized a heterozygous pathogenic missense in *RRM2B* (c.328C>T, p.Arg110Cys). But since all other variants in the gene were benign (c.207C>T, p.Val69=) or intronic, the gene was not followed up as potentially disease-causing. As individual #126118 remained inconclusive after WES, RNA-seq was performed. OUTRIDER reported *RRM2B* as expression outlier with a 42% reduction, which matched the expectations due to the heterozygous pathogenic missense variant (Fig. 5.3a-b). But interestingly, the MAE analysis found two hits in *RRM2B* including the pathogenic variant (Fig. 5.3c) meaning that the reduction in gene expression can not be linked to the pathogenic variant and hence is caused by other means. Unfortunately, WGS analysis did not reveal any additional potentially disease-causing variants and also the aberrant splicing analysis was negative. To investigate the pathogenicity of the missense variant, proteomics was performed. This confirmed the loss of function as only 28% of RRM2B was detected harboring the pathogenic amino acid change. The nuclear encoded *RRM2B* is important in the DNA synthesis and hence, RRM2B-deficiency (OMIM #604712) is linked with the mitochondrial DNA depletion syndrome[Bourdon et al., 2007] but also linked to multiple

distinct clinical characteristics[Pitceathly et al., 2012] matching some of the presentations of individual #126118. Taking together the RNA-seq based and proteomics based evidences, the MAE of the pathogenic missense variant in *RRM2B* is the molecular cause of the disease, but as the evidence for the loss of expression of the second allele is still unclear, further investigations are needed to fully understand the pathomechanism.

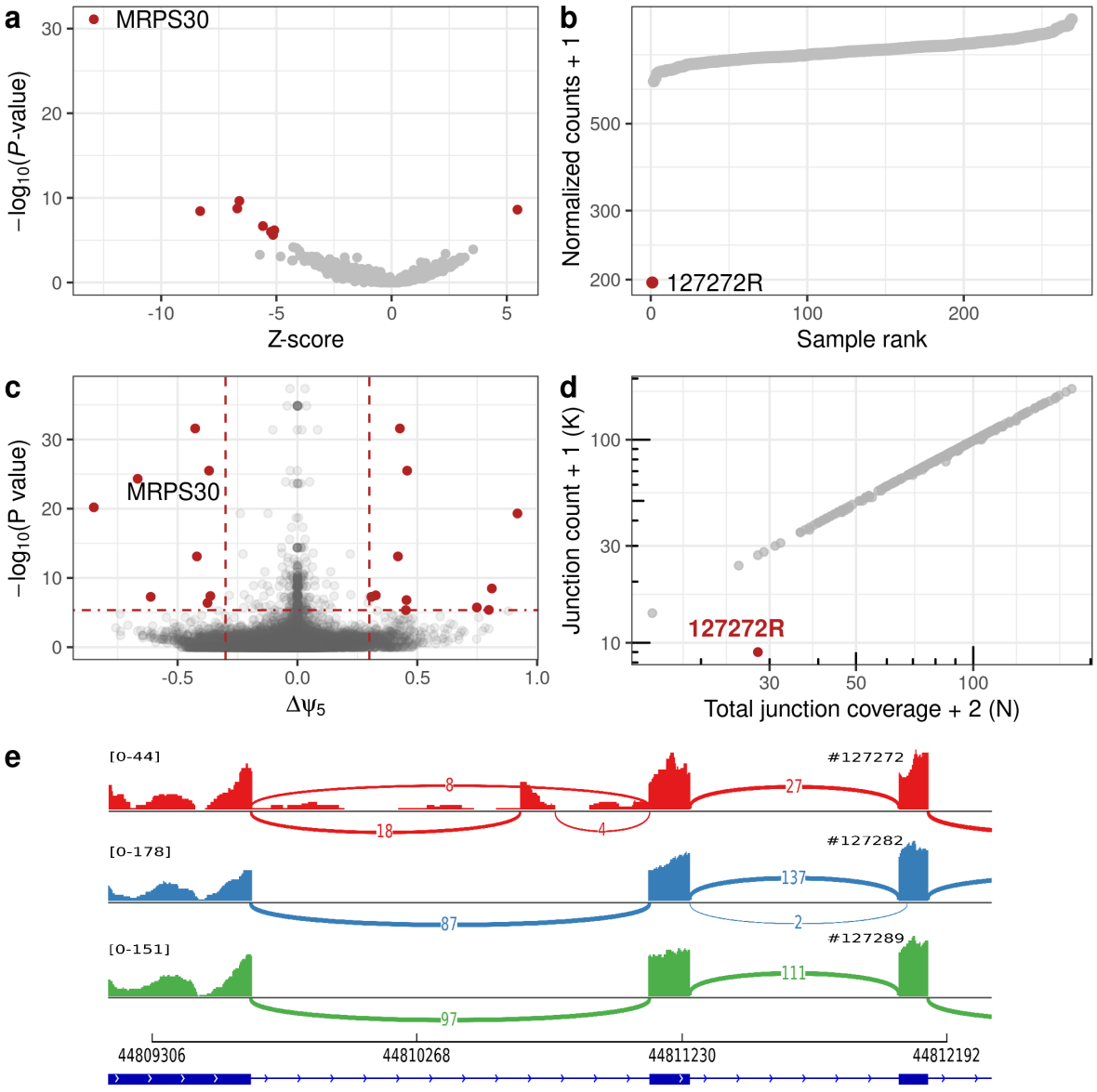

Figure 5.2: Expression of a pathogenic cryptic exon in *MRPS30*. **(a)** Gene-level significance (-$\log_{10}(P)$) versus effect (z score) for individual #127272 for gene expression using OUTRIDER. Each dot represents a gene and red dots indicate genome-wide significance. **(b)** Normalized gene expression versus sample rank for *MRPS30*. **(c)** Gene-level significance (-$\log_{10}(P)$) versus effect ($\Delta\psi_5$) for individual #127272 for alternative acceptor usage using FRASER. The red dashed lines represent the genome-wide significance cutoff (horizontal) and the effect size cutoff (vertical). **(d)** Number of split reads spanning the canonical intron (y-axis) against the total number of split reads at the donor site (x-axis) of *MRPS30*. **(e)** Sashimi plot of the cryptic exon event in RNA-seq samples of individual #127272 (red) and two non-affected individuals (blue and green). The bottom depicts the RefSeq annotation.

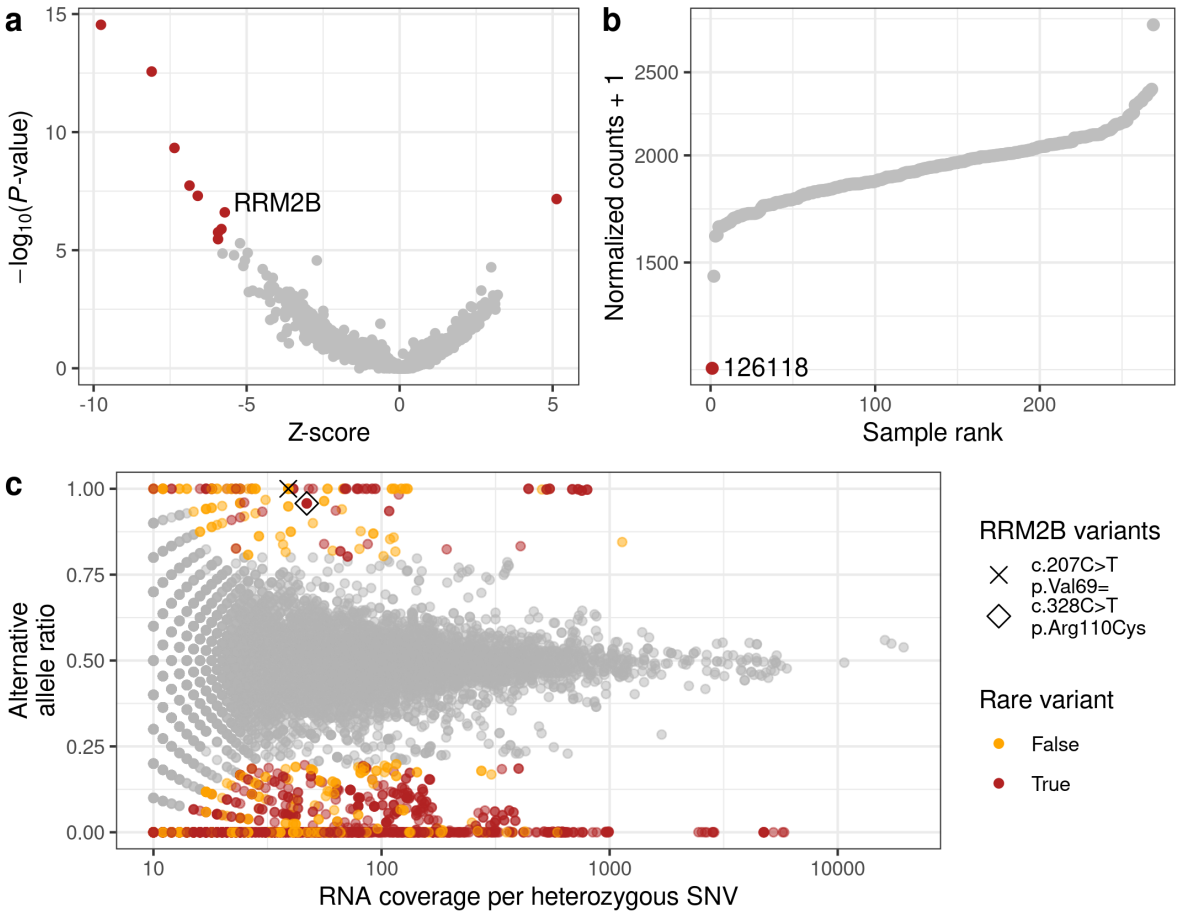

Figure 5.3: Mono-allelic expression of rare missense variant in *RRM2B*. **(a)** Gene-level significance (-$\log_{10}(P)$) versus effect (z score) for individual #126118 for gene expression using OUTRIDER. Each dot represents a gene and red dots indicate genome-wide significance. **(b)** Normalized gene expression versus sample rank for *RRM2B*. **(c)** Alternative allele ratio (y-axis) versus total coverage of heterozygous SNVs (x-axis) for the sample #126118. Significant MAE in common and rare variants are shown in orange and red, respectively. MAE variants in the disease-causing gene are highlighted by different shapes.

6 Conclusion

Over the past decade, NGS has revolutionized the clinical diagnostics of rare diseases. Rapid technology development and drastically reduced sequencing costs made WES and eventually WGS the entry point for molecular diagnosis of rare diseases. The clinical implementation of WES and WGS yields a diagnostic rate between 25-50% depending on the disease.[Neveling et al., 2013; Wortmann et al., 2015; Wright et al., 2018b; Retterer et al., 2016] To increase this diagnostic rate and to solve the remaining cases, the community is turning to the *next* omics level, namely the RNA.[Cummings et al., 2017; Kremer et al., 2017; Gonorazky et al., 2019; Frésard et al., 2019; Murdock et al., 2021] This new development and endeavor allowed me, in this thesis and beyond, to systematically investigate the utility and complementarity of RNA-seq in rare disease diagnostics. Further in this thesis, I have developed statistical methods to increase sensitivity, specificity, and robustness in detecting aberrant gene expression and splicing events using RNA-seq data to support molecular diagnosis.

Establishment of RNA-sequencing-based rare disease diagnostics In the first part of this work, we aimed to demonstrate the utility of RNA-seq and its complementarity to WGS in rare disease diagnostics. By adapting methods developed for differential expression[Love et al., 2014] and splicing analysis,[Li et al., 2018] we were able to provide a molecular diagnosis for 10% (5 out of 48) of WES inconclusive cases. With the significance-based tests for aberrant gene expression and splicing events and MAE of rare variants, we limited the number of outlier events per individual to a median of 1, 5, and 6, respectively. Such a small number of candidate genes per sample is important to not overwhelm and importantly not distract clinicians with too many non-disease-relevant hits while manually inspecting the results. To minimize the number of covariates and confounders and hence the extensive normalization of the data later, the same sample preparation and sequencing protocols were used. Still, three covariates were needed to normalize the data to an acceptable level including sex, biopsy site, and an unknown probably technical bias. Despite this knowledge of existing covariates, the splicing analysis was performed without normalization as LeafCutter did not provide the functionality at the time.[Li et al., 2018] As more data will be generated over time and across hospitals,[Gahl et al., 2015; Thompson et al., 2014; Frésard et al., 2019] including healthy cohorts such as GTEx,[The GTEx Consortium et al., 2015] the need to control for confounders became evident and led us to the development of more robust statistical models for aberrant gene expression and splicing events.

In the case of MAE, we did not control for any biases as potential covariates are reduced to a bare minimum. This is due to the test within a single biological sample between the two alleles. The mapping bias towards reference alleles can be addressed

by mapping in a variant aware fashion using WASP.[van de Geijn et al., 2015] Another approach to test for MAE is to estimate the allelic imbalance per gene across a control population and then test each sample for significant divergence from it as done by Mohammadi et al. [2019]. This enables individual transcriptome comparison to previously generated reference data, but with the caveat of loosing many genes where no estimates could be fitted due to lack of expression or non-existing heterozygous variants in the reference cohort. Therefore, a combination of the proposed analysis using the expression variation dosage outlier test (ANEVA-DOT) for genes where estimates can be established while using the DESeq2-based[Love et al., 2014] NB test for the remaining ones.

Interestingly, I observed many aberrant splicing defects in our pilot dataset causing loss of function that were ultimately leading to a molecular diagnosis in 4 cases. This finding was in line with the results from Cummings et al. [2017] were the majority of disease-causing events were linked to splicing defects found in individuals affected by a primary muscle disorder using muscle biopsies. This confirms the important role of mis-splicing in Mendelian[Sibley et al., 2016; Abramowicz and Gos, 2018] as well as in common diseases.[Li et al., 2016; Scotti and Swanson, 2015] To detect these events DNA sequencing is not enough as predicting the effect of variants on splicing is still challenging,[Jian et al., 2014] especially for deep intronic variants.[Jaganathan et al., 2019] An example of this is the splicing defect in *TIMMDC1* and *MRPS30*, where a single deep intronic variant caused aberrant expression of a cryptic exon that was neither detected by WES nor prioritized by prediction models.[Jaganathan et al., 2019; Cheng et al., 2019] Here RNA-seq was crucial to pin-point the disease-causing event as well as the underlying variant. By investigating these cryptic exon events, I discovered that over 70% of such events originated from splice sites expressed below 1% in the population. As these weak splice sites are more likely to be activated to create a cryptic exon by a single variant as the data suggest, GTEx could be used to identify these sites. Such a weak splicing map can be used to improve variant filtering but also be used to train splicing models like MMSplice and SpliceAI even in a tissue-specific manner.[Cheng et al., 2019; Jaganathan et al., 2019]

It is known that splicing and gene expression is tissue specific.[GTEx Consortium, 2017; Yeo et al., 2004a; Castle et al., 2008] This naturally raises the question if clinically accessible but unaffected tissue can be used to detect a disease-causing event.[Aicher et al., 2020] In this study patient derived dermal fibroblast cell lines were used as a proxy instead of affected tissue for mitochondrial diseases as most mitochondrial genes are expressed in fibroblasts.[Vafai and Mootha, 2012; Yépez et al., 2021a] In contrast, Cummings et al. [2017] and Gonorazky et al. [2019] demonstrated that muscle biopsies are needed in neuromuscular diseases, as disease relevant genes would otherwise be missed due to missing expression. One way to use clinical accessible tissues is to transdifferentiate patient derived fibroblasts into the target tissue as proven effective for neuromuscular diseases by Gonorazky et al. [2019]. In fact, it can be argued that using unaffected tissue over affected tissue may actually be advantageous because regulatory effects are limited to other genes as long as the gene of interest is expressed, making causal defects more likely to be outliers.

Over the time, multiple studies showed the advantage of using RNA-seq to diagnose WES or WGS inconclusive cases.[Cummings et al., 2017; Gonorazky et al., 2019; Frésard et al., 2019; Murdock et al., 2021; Yépez et al., 2021a] While all studies had the same aim and concluded that RNA-seq was beneficial, the approaches and especially the way aberrant events were detected differed. Rather than using the cohort itself as a control, as done in this study, Cummings et al. [2017] compared expression and splicing profiles with tissue-matched control samples from GTEx. Specifically, gene expression outliers were identified using a cutoff on z scores ($|z| > 3$) calculated on the log-transformed RPKM values without controlling for covariations. This approach was adapted by Gonorazky et al. [2019] with a lower cutoff ($|z| \geq 1.5$). While the first study did not identified any gene expression outliers claiming that the analysis was statistically underpowered with more than 180 samples,[Cummings et al., 2017] the latter study was able to diagnose patients with this approach.[Gonorazky et al., 2019] To account for covariates in the data, Frésard et al. [2019] controlled the RNA-seq data using PCA before applying a z score cutoff.[Frésard et al., 2019] This was important, not only because the samples were sequenced in different centers across multiple batches and projects but also to control for hidden factors. This was similar to the approach by Li et al. [2017] in the context of the general impact of rare variants on gene expression, where the data was controlled for genotype, sex, and hidden confounders found by PEER.[Stegle et al., 2012] Compared to the one-against-the-rest DESeq2 approach, all methods used some log transformation of the raw read counts while assessing the outlier status based on z scores instead of significance.

Comparing the aberrant splicing detection approaches across the studies, revealed major differences in the methodology.[Cummings et al., 2017; Gonorazky et al., 2019; Frésard et al., 2019; Murdock et al., 2021; Yépez et al., 2021a] Nevertheless, all found disease-causing aberrant splicing events. Cummings et al. [2017] and Gonorazky et al. [2019] used a similar approach based on cutoffs on absolute and relative RNA-seq split read counts. Because the cutoffs were chosen manually and not evaluated in other scenarios, it is questionable whether they generalize across cohorts and diseases. In particular, filtering for introns not expressed or expressed in fewer than five unaffected samples would not detect any cryptic exons arising from weak splice sites such as *TIMMDC1* or *MRPS30*. Frésard et al. [2019] adapted the z score approach from the gene expression analysis by using PCA controlled splice ratios to compute z scores. These methods are in contrast with the approach I developed by adapting LeafCutter to the outlier test, as they do not assess the significance per event or provide data driven cutoffs. Following this idea, LeafCutter was recently extended to test properly for aberrant splicing based on the Dirichlet-multinomial distribution without a thorough evaluation against alternative methods.[Jenkinson et al., 2020]

Development and benchmarking of robust gene expression and splicing outlier detection methods All these studies including the pilot study I described here, had at least two out of the four following limitations: (i) no assessment of the significance of outlier events, (ii) no controlling for known or unknown confounders, (iii) no usage of

appropriate count distributions, and (iv) no thorough evaluation of the methods using simulated or experimental data. These limitations led us to the development of the two software packages OUTRIDER and FRASER and to the new benchmark strategies for gene expression and splicing outlier detection in RNA-seq, respectively. Both methods are using an autoencoder to control in an automated fashion for confounding effects in the RNA-seq data by providing expression or splice ratio estimates. Using these estimates appropriate count distributions, namely NB and BB, are fitted per feature across samples accounting for overdispersion in RNA-seq count data. Outliers are then identified as read counts or split read counts that significantly deviate from the fitted distribution. The number of detected confounders to control for, is estimated by optimization of the model's ability to recall corrupted counts using a denoising autoencoder scheme. With the newly developed benchmark strategies, I demonstrated that each method outperformed alternative methods in recalling simulated outliers and pathogenic events in the Kremer dataset. Calling aberrant expression events with OUTRIDER and FRASER yielded in higher enrichments for rare moderate and high impact variants as well as for rare splice effecting variants across the GTEx tissues.[The GTEx Consortium et al., 2015] Both methods are developed as open source R package and made available through Bioconductor.[Team, 2021; Huber et al., 2015] The packaging of the software together with the comprehensive vignettes and user documentation, makes the adaptation of the tools as easy as possible for the end-user and a full analysis on RNA-seq data can be done in only a few lines of code.

Controlling for confounders in RNA-seq data can be achieved in multiple ways. In the DESeq2-based expression outlier detection approach presented in this study, manually identified covariates were regressed out. Another way is to use PEER[Stegle et al., 2012] in conjunction with known covariates like sex and genotype [Li et al., 2017; Pala et al., 2017; Lappalainen et al., 2013] or PCA[Pickrell et al., 2010; Frésard et al., 2019] to control gene expression. Not only gene expression but also splicing metrics need to be controlled, as demonstrated by Frésard et al. [2019] using a PCA, despite the initial misconception that splice ratios have implicit normalization.[Li et al., 2018] All of these approaches transform the data into the log space or into ratios assuming a log-normal or normal distribution, respectively. This is suboptimal when working with overdispersed count data. Indeed, using PCA or PEER on simulated data instead of the autoencoder, increased the error of inferring the expectations especially for genes with low or high expression values. Even though it was not much, for the splice metrics performance increased when a BB loss function was used for the decoder. Instead of changing the methodology one can also change the architecture of the autoencoder. I did not explore more complex architectures with multiple layers for the autoencoder to capture nonlinear relationships as the single layer autoencoder already removed almost all existing correlations. This is in line with the work by Way and Greene el al., who demonstrated that an autoencoder with a single layer for the encoder and decoder was enough to learn biologically relevant features from gene expression data in cancer.[Way and Greene, 2018]

In addition to the modelling of the covariations in the data, I demonstrated the benefits of using P values over z scores to identify aberrant data points through extensive

benchmarking. Until now, aberrant gene expression and splicing was detected by z score cutoffs only.[Cummings et al., 2017; Frésard et al., 2019; Gonorazky et al., 2019; Li et al., 2017] Only recently with the development of LeafCutterMD[Jenkinson et al., 2020] and SPOT,[Ferraro et al., 2020] aberrant splicing events were detected via statistical means. The advantage of using P values is two fold. First, P values can be corrected for multiple testing,[Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001] which is important as more than 10,000 test per sample are performed. Second, by using z scores the input data is transformed assuming a normal distribution resulting in overall lower precision. Using P values allows to apply appropriate count distributions, which is especially important for features not well expressed.

In differential expression analysis, two conditions are usually compared, each with at least three replicates.[Love et al., 2014; Zhou et al., 2014; Li et al., 2018] This is quite different to the experimental setup of outlier detection in rare diseases where no replicates are performed. It is assumed that the disease-causing event has such an effect on the expression that no replicates are needed to identify it. As the past showed, no technical or biological replicates were performed in the context of rare disease diagnostic.[Cummings et al., 2017; Frésard et al., 2019; Gonorazky et al., 2019; Murdock et al., 2021; Yépez et al., 2021a] OUTRIDER and FRASER as well as all the other methods assume that disease-causing events are ultra rare and hence are not replicated in the data, because this could render the event undetectable. In this context, replicates can be interpreted in two ways: (i) technical/biological where two samples are sequenced for a single individual or (ii) event-based were two individuals have the same underlying genetic cause resulting in the same aberrant event. The event-based replicates can occur, especially when families with multiple affected children are analyzed. To not lower the statistical power, the autoencoder implementation allows for exclusion in the modeling step of such real or pseudo-replicates. While all samples are tested independently for outlier events, the user can leverage such replicates by pairing the them using Fisher's method of combining P values[Fisher, 1970] by assuming independence of the read counts conditioned on the expectation predicted by the autoencoder. Alternatively, methods from studies where family structures were present can be leveraged.[Pala et al., 2017; Li et al., 2014] But to evaluate the overall benefit of using replicates, more investigations are needed.

It is known that splicing is a very sophisticated process that can generate tens of isoforms per gene with complex splicing patterns even in the same tissue.[Mortazavi et al., 2008; Vaquero-Garcia et al., 2016] To keep the model simple, as well as to simplify the interpretation of the results, I developed FRASER using the intron-centric splicing metrics which are defined at the level of individual splice sites. Using the intron-centric percent spliced in values (ψ_5 and ψ_3)[Pervouchine et al., 2013] has also the advantage that they can be computed without a gene annotation. This is in contrast to existing differential splicing analysis tools that uses exon-centric ψ values or exon and intron quantifications to model alternative splicing and intron retention.[Anders et al., 2012; Shen et al., 2014; Trincado et al., 2018] Using the ψ_5 and ψ_3 values also enabled me directly to use the splicing efficiency metrics (θ_5 and θ_3)[Pervouchine et al., 2013] with the same autoencoder architecture since they are computed in the same way as count ratios. Alternatively, aberrant splicing can be modelled on the gene level using a Dirichlet-

multinomial distribution.[Li et al., 2018; Jenkinson et al., 2020; Ferraro et al., 2020] Theoretically, this should increase sensitivity over a BB model as all split reads across the gene contribute to the test statistic. Nevertheless, FRASER with its splice-site-level statistics outperformed the gene-level approaches. In addition, the gene-level approaches have two drawbacks. First, to assign each intron to a gene, a complex clustering must be performed or a known gene model is required a priori.[Li et al., 2018] Second, since only split reads are considered, by design the Dirchilet-multinomial approach is blind to intron retention, which is proven to be important in clinical diagnostics.

In summary, by developing dedicated statistical models to detect aberrant expression events in RNA-seq, I directly contributed to the diagnoses of more than 40 individuals over the last years. This has likely not only improved the quality of life of each patient but also demonstrated how RNA-seq can be used to support rare disease diagnostics but also variant interpretation in general by providing complementary information through aberrant expression event detection. With the falling costs for RNA-seq, the increased implementation of RNA-seq-based diagnostics, and the advantage of OUTRIDER and FRASER over alternative approaches, I foresee that they will become important tools in the field of rare diseases.

6.1 Outlook

This work began with a systematic evaluation of the utility of RNA-seq in clinical rare disease diagnostics. This exposed the lack of dedicated statistical models to detect outlier events in RNA-seq, which I addressed by developing OUTRIDER and FRASER. Even though, these methods outperformed alternative methods in the benchmarks, there is room for further improvement and possible adaptations to new fields and problems.

Improving the autoencoder model by incorporating known covariates and robust estimates The autoencoder implementation does not allow the usage of known covariates. As long as the data is homogeneous, e.g from the same tissue or produced through the same pipeline, the autoencoder is capable of remove any covariations. But when combining more heterogeneous datasets like the Kremer and GTEx dataset, some structure is left even after the autencoder correction.[Yépez et al., 2021b] Adding known covariates like sequencing center and protocol, sex, and age to the autoencoder model could be achieved by adding them along with the latent factors before the decoder layer.

The dispersion parameters are fitted on the full data including the outlier, if present. As outlier data points can have high impacts on the fitted parameters, it could be investigated if a weighting scheme as implemented in the edgeR package[Zhou et al., 2014] can further improve the results. This of course introduces some sort of circularity and hence has to be evaluated.

From single- to multi-omics outlier detection This study focuses primarily on three aspects of RNA-seq sequencing to guide and support the diagnostics of rare disease:

aberrant expression and splicing and MAE. Through the rapid advances in technology, new avenues are arising each with its own advantages and drawbacks. The undiagnosed disease network (UDN) for example, used metabolomics and lipidomics to prioritize and identify disease-causing events.[Webb-Robertson et al., 2020] Multiple studies used proteomics to validate candidate genes,[Kremer et al., 2017; Lake et al., 2017; Borna et al., 2019] but it can also be used directly to identify the disease-causing event.[Kopajtich et al., 2021] Instead of looking at the population of cells as a whole, one can even go down to the level of a single cell.[Regev et al., 2017] Nomura et al. [2018] used single-cell transcriptomics and epigenomics to identify functional signature in cardiac hypertrophy. In another study,[1] transcriptomics and proteomics on the single-cell-level are used to understand muscle dystrophy.

We already adopted successfully the autoencoder approach to proteomics data.[Kopajtich et al., 2021] But more work is needed to adapt it to other omics and single cell data and to go truly multi-omics by jointly modeling the different omics datasets. Argelaguet et al. [2018, 2020] already demonstrated the possibility and actual benefits of learning across omics and single-cell datasets. This is promising as it may increase sensitivity but also interpratability, as the effect of a disease-causing variant can be detected at multiple levels. For example, downregulation of *ALDH18A1* expression leads to aberrant changes in metabolites of the proline synthesis or the splicing defect in *TIMMDC1* leads to downregulation of mitochondrial complex I subunits at the protein level.

Community outreach Just developing methods and proving their superior over alternative methods is not enough to improve rare disease diagnostics and hence the patient's quality of live. It needs outreach, accessibility, training, documentation, support, and foremost a user community. This is an ongoing process. By making the source code open access and integrating it into the Bioconductor[Huber et al., 2015] and bioconda[Grüning et al., 2018] ecosystem with detailed vignettes, we made it accessible and provide documentation. The methods developed in this thesis are even integrated into DROP a end-to-end workflow for aberrant event detection in RNA-seq to ease the implementation of the tools in diagnostics.[Yépez et al., 2021b] With the growing interest in using RNA-seq in diagnostics, I organized jointly with colleagues from the field interactive workshops to bioinformaticians and clinicians to provide best practices but also to start the conversions with the community on how to use RNA-seq for rare disease diagnostics.[2,3,4] Such events are important to foster community where best practices, ideas, and challenges are communicated and discussed. They also help to widen the user base and to create new collaborations that ultimately help patients. Such an example is the study by Murdock et al. [2021] that resulted in a collaboration and the adaptation of the DROP pipeline after the attendance of the interactive workshop at ASHG 2019.

[1] MYOCITY (Ref: EJPRD19-118), a multidimensional single-cell approach to understand muscle dystrophy

[2] Interactive workshop at ASHG 2019: RNA-seq for Mendelian Diseases Diagnostics (Session ID: 710)

[3] Interactive workshop at ASHG 2020: RNA-seq for Mendelian Diseases Diagnostics (Session ID: 133)

[4] Interactive workshop at eMed 2020: Detecting aberrant expression in RNA-seq

Improving diagnostics by sharing NGS-based data at large scale Sharing of knowledge and data is the key to success in rare diseases.[Boycott et al., 2017] Hence multiple local communities and databases were connected to *match* individuals based on genotype or phenotype in a global scale through the Matchmaker Exchange platform to build evidence for causality.[Philippakis et al., 2015; Brookes and Robinson, 2015] This effort amplified and accelerated disease-gene discovery and include projects such as LOVD,[Fokkema et al., 2011] DECIPHER,[Swaminathan et al., 2012] RD-connect,[Thompson et al., 2014] and UDN,[Brownstein et al., 2015] just to name a few. Another evidence of this is the great success of the global variant frequency databases ExAC and gnomAD that contains thousands of WES and WGS samples collected across the world and is now used as the gold standard to filter out non-disease-causing rare variants.[Lek et al., 2016; Karczewski et al., 2017, 2020] Therefore, over the last years multiple large scale national and international projects started to take the challenge of improving rare disease diagnostics by sequencing DNA and RNA at large while sharing the data with the research community, which includes the 100k genomes project,[Genomics England, 2017] All of Us,[All of Us, 2019] GHGA,[5] genomDE,[6] and MEGA.[7] All of these projects include phenotypic data to varying degrees, which opens up new research opportunities, but also unique challenges. In the lens of RNA-seq-based diagnostics, having more data allows better distribution estimates resulting in higher sensitivity in expression outlier detection. Registering the disease-causing event in the case of molecular diagnosis, will turn such efforts into great benchmark and training sets with thousands of validated data points. Similar to ClinVar at the variant level,[Landrum et al., 2018] but for variants directly effecting the transcriptome coupled with the original RNA-seq data. Currently outlier detection methods are developed and benchmarked on GTEx and in-house datasets with only a hand full of known disease causing events which can lead to overfitting to a particular use case. Hence, sharing thousands of WGS and matching RNA-seq samples across a variety of rare diseases will boost the development of such methods.

Increasing the available cohort size from hundreds to thousands or even to millions brings unique challenges for the data analysis, applied methods, visualization, workflows, privacy, and the underlying IT infrastructure. The cloud is in this case not always the right answer, as long-term costs and privacy concerns arise. Therefore, federated systems are implemented to allow and account for country- and state-specific laws and regulations. Incorporating data across such a federated system and from multiple sources at this scale requires appropriate and sophisticated quality control and quality metrics next to well maintained metadata to allow researchers to filter datasets for downstream analysis according to their research needs. On top of this, the methods developed for outlier detection in RNA-seq data have never been performed with more than 1000 samples, let alone applied in a federated way. Running them in such a federated system on thousands of samples require adaptations and further development.

[5] Deutsches Humangenom-Phenomarchiv – GHGA (DFG Proj: 441914366)
[6] Die deutsche Genom-Initiative – genomDE
[7] 1+ Million European Genomes Initiative – MEGA

Despite the unique challenges, these large-scale national and international efforts will advance rare disease diagnostics by increasing the diagnostic rate through data sharing. They will provide an unmatched treasure trove of biomedical data for the rare disease and research community, and will therefore transform the way rare disease diagnostics, as well as basic research, are conducted by bringing them closer together.

A Appendix

A.1 Webresources

GTEx Portal, https://www.gtexportal.org/home

OMIM, http://www.omim.org

OUTRIDER package, http://bioconductor.org/packages/OUTRIDER/

OUTRIDER analysis pipeline, https://github.com/gagneurlab/OUTRIDER-analysis

FRASER package, http://bioconductor.org/packages/FRASER/

FRASER analysis pipeline, https://github.com/gagneurlab/FRASER-analysis

DROP source, http://github.com/gagneurlab/drop

A.2 Appendix: Supplemental Figures

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Figure A.1: ACMG Evidence Framework The following chart organizes each of the criteria by the type of evidence as well as the strength of the criteria for a benign (left side) or pathogenic (right side) assertion. Evidence code descriptions can be found in Richards et al. [2015] Table 3 and 4. Abbreviations: BS, benign strong; BP, benign supporting; FH, family history; LOF, loss-of-function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong Taken from Richards et al. [2015].

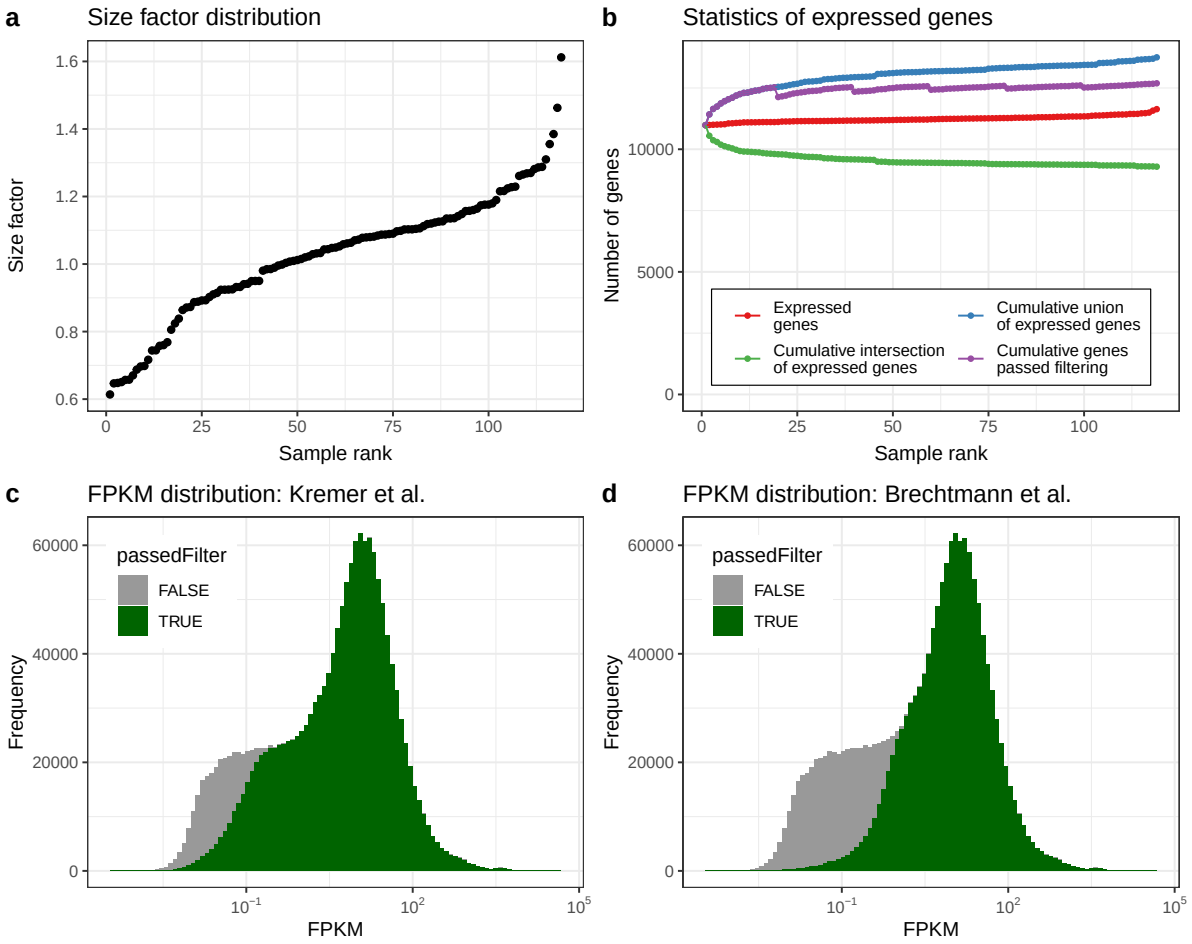

Figure A.2: QC and filtering statistics for the Kremer dataset (a) The estimated size factor plotted against its rank. **(b)** Number of expressed genes cumulative across all samples. Colors represent the union of all detected genes (blue), genes that passed the raw read count filter described in Section 2.2.1 as a group (violet), genes that are expressed in each sample (red), and the intersection of expressed genes (green). **(c)** Histogram of the FPKM values grouped according to the filter status described in Section 2.2.1. Green indicates the genes that passed the filter and gray those that were filtered out. **(d)** Same as c, but according to the filtering steps as described in Section 3.2.

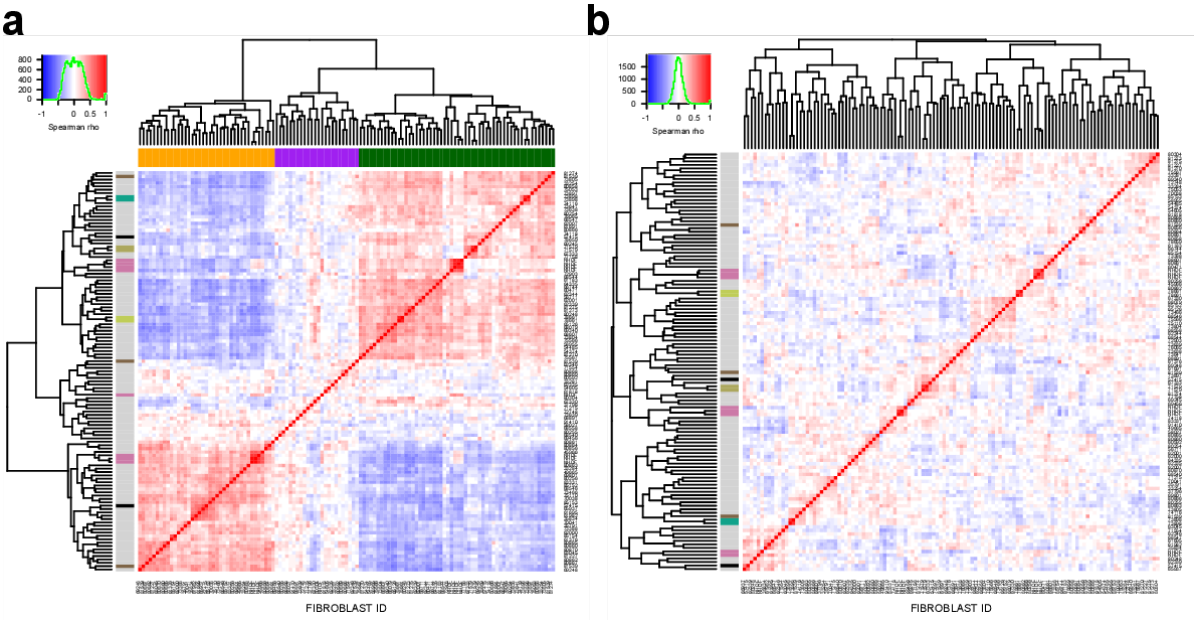

Figure A.3: RNA-seq normalization in Kremer et al. [2017]. **(a)** Spearman correlation heat map of size-factor normalized gene expression between all fibroblasts (n=119) including biological replicates (left side color code). The dendrogram represents the sample-wise hierarchical clustering. The color code on the top depicts the top three clusters. The color key of the spearman rho value (top left) includes a histogram based on the values (green line). **(b)** Same as a after normalization for the technical variation, sex variation, and four HOX gene groups. Adapted from Kremer et al. [2017].

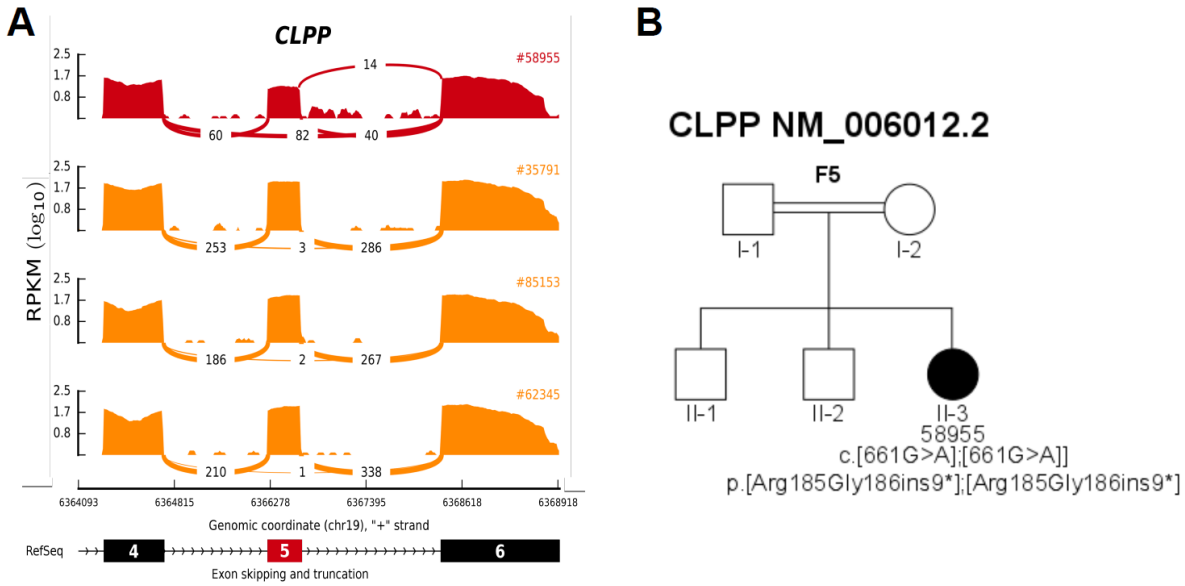

Figure A.4: Exon skipping in CLPP (A) CLPP Sashimi plot of exon skipping and truncation events in affected and unaffected fibroblasts (red and orange, respectively). The RNA coverage is given as the $\log_{10}$ RPKM-value and the number of split reads spanning the given intron is indicated on the exon-connecting lines. At the bottom the gene model of the RefSeq annotation is depicted. The aberrantly spliced exon is colored in red. **(B)** Pedigree of the family with mutations in *CLPP* showing the mutation status. Adapted from Kremer et al. [2017].

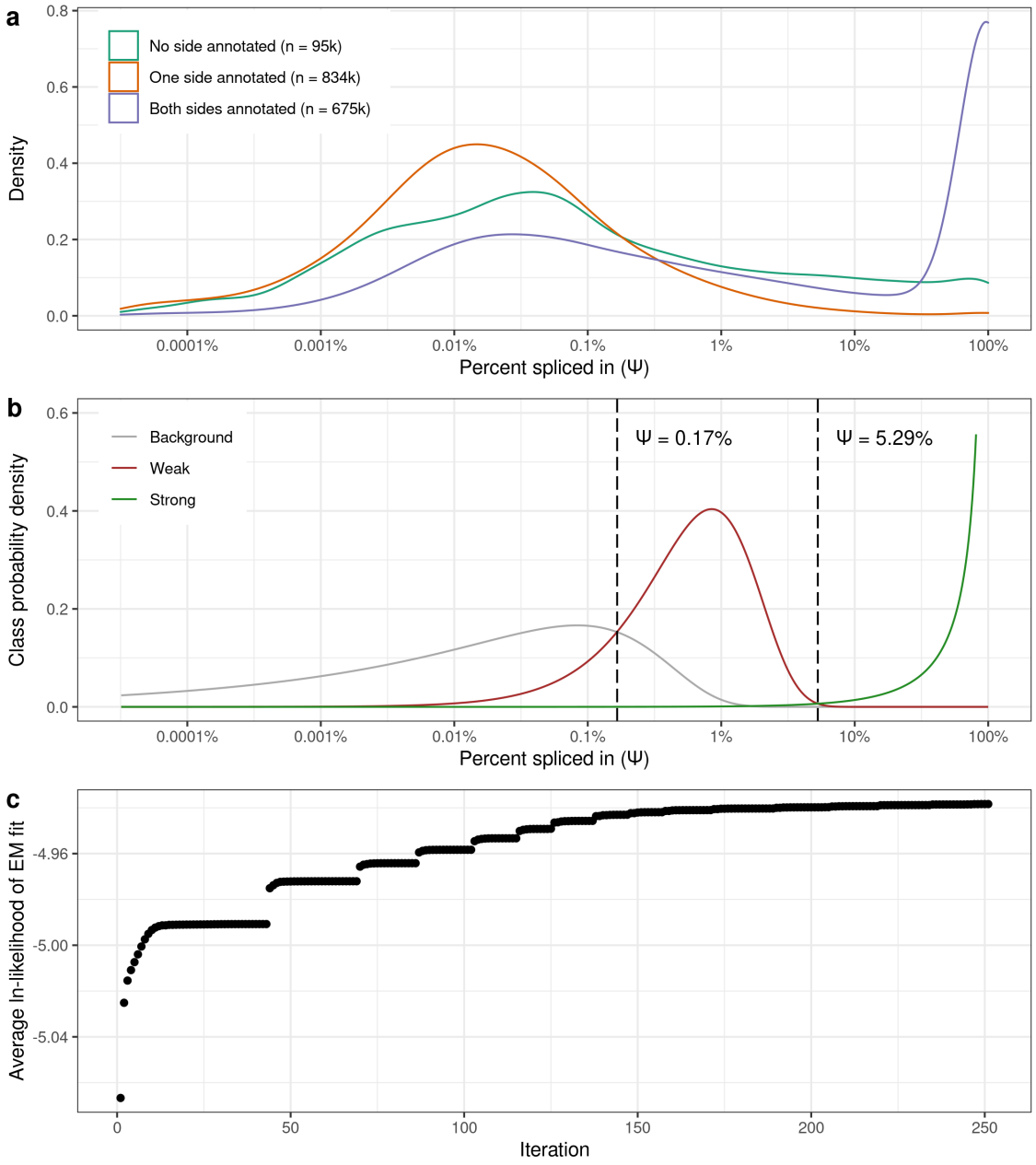

Figure A.5: Percent spliced in distributions. (a) The densities of genome-wide ψ_5 and ψ_3 (percent spliced in) values grouped by their GENCODE annotation status: both sites of the junction are annotated (green), only one site of the junction is annotated (orange), and no site of the junction is annotated (blue). **(b)** The splice class model fitted by expectation maximization (EM) based on the GENCODE annotation status. Each line represents the probability density of belonging to a splice class given a ψ value. The dash lines depict the lower and upper boundary of the weak splicing class **(c)** The convergence of the EM algorithm. Each point represents the average ln likelihood of the EM fit after a given iteration cycle (n=250). Adapted from Kremer et al. [2017].

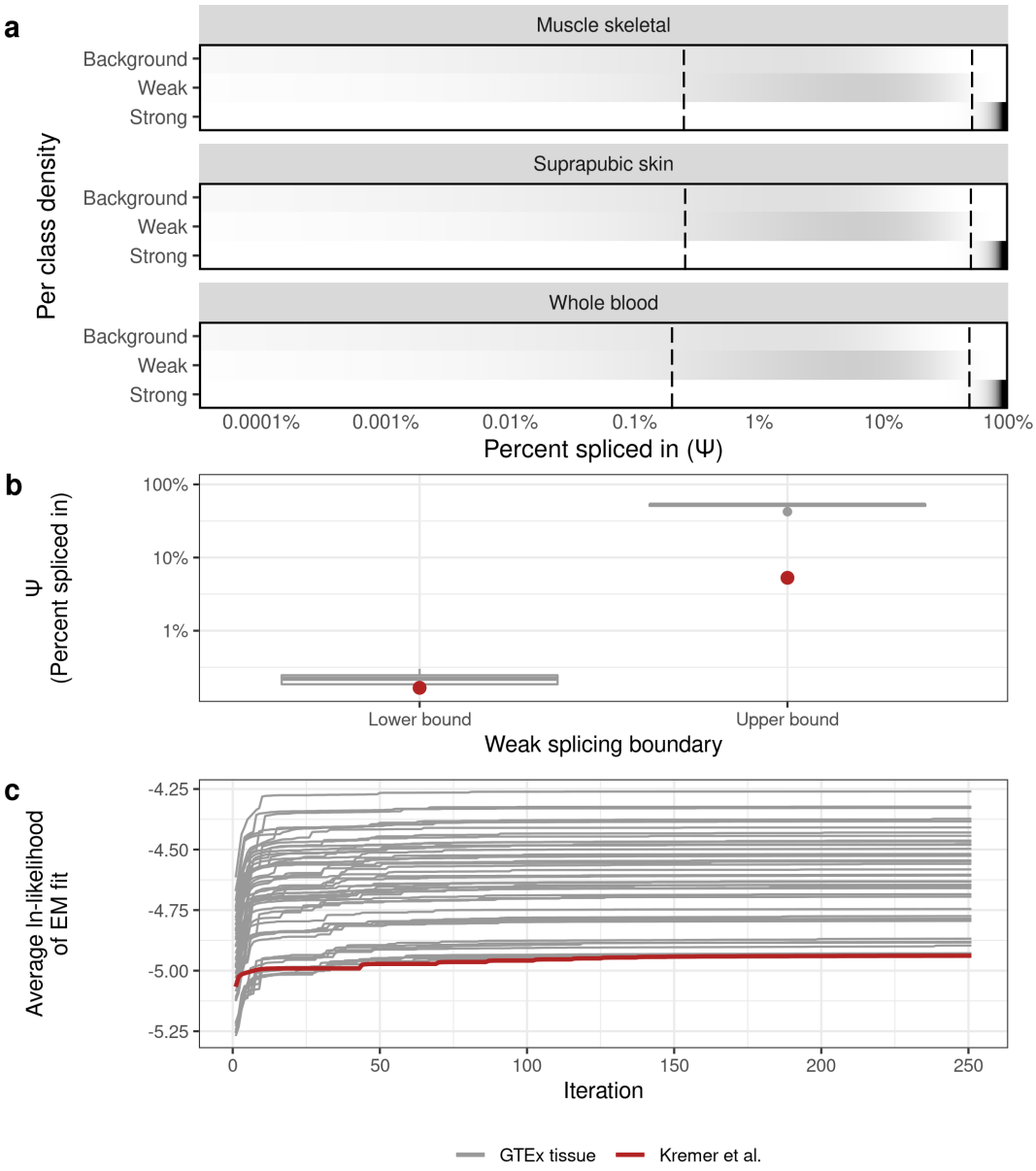

Figure A.6: Weak splicing in GTEx. **(a)** The density of the splicing class probability for *background*, *weak* and *strong* are shown, respectively. The darker the gray the higher the density being the given class. The dashed lines depict the lower and upper boundaries of the weak splicing class. It is faceted by the choosen representative GTEx tissues: muscle skeleton, suprapubic skin, and whole blood. **(b)** Lower and upper boundary distributions of the weak splicing class across all GTEx tissues. The red point depicts the boundaries observed by Kremer et al. [2017]. **(c)** The convergence of the EM algorithm for each GTEx tissue. Each line represents the course of the average ln likelihood of the EM fit in a given GTEx tissue over all iteration cycles (n=250). The red line depicts the average ln likelihood presented by Kremer et al. [2017].

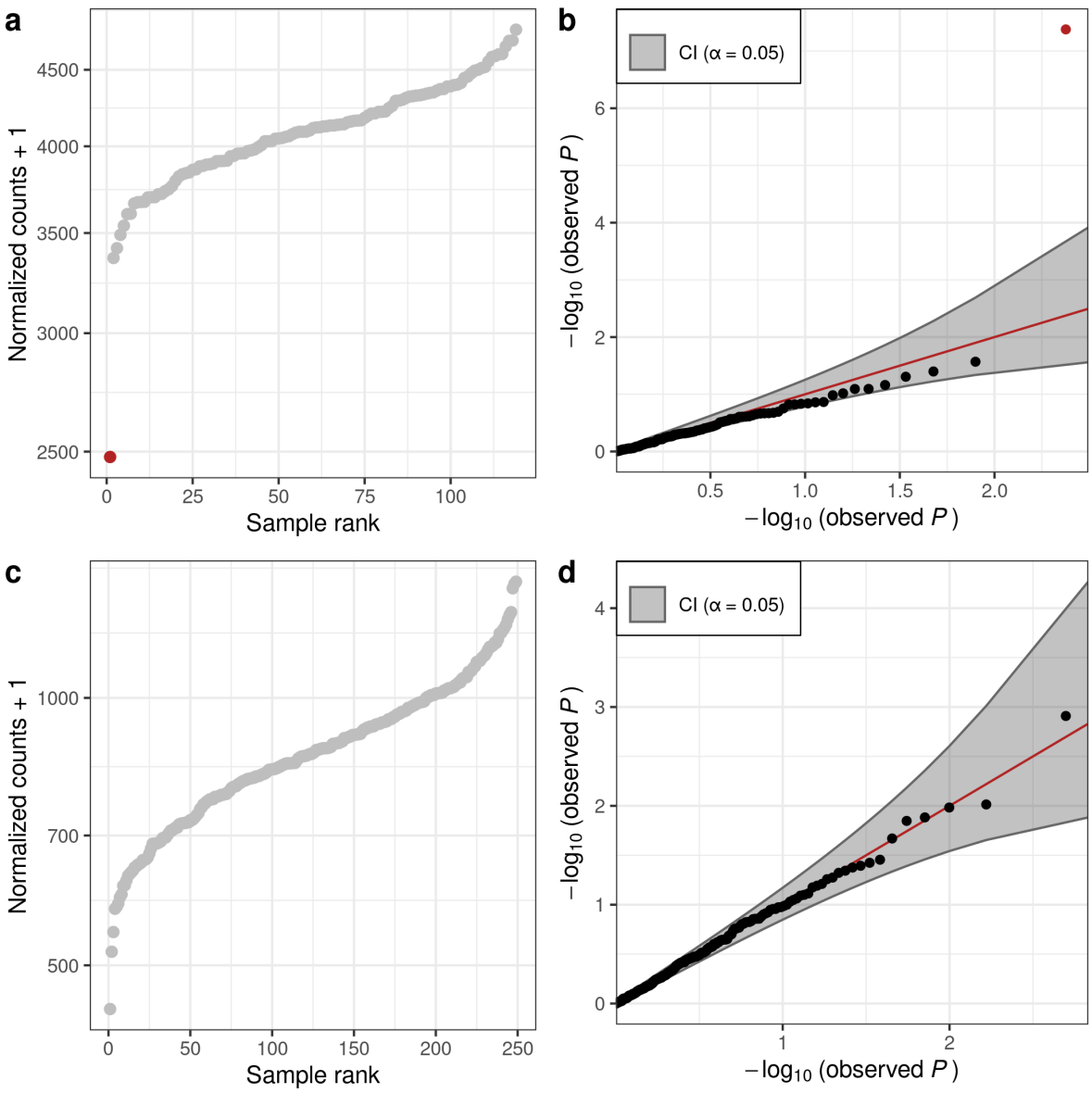

Figure A.7: Using the NB distribution for significance assessment (a) Normalized RNA-seq read counts plotted against their rank for *ALDH18A1* in the Kremer dataset. Expression outliers are shown in red (FDR < 0.05). **(b)** Quantile-quantile plots of observed P values against expected P values with 95% confidence bands for data in a. **(c)** Same as a but for *CDCA7* in the GTEx suprapubic skin tissue. **(d)** Same as b but for the data in c.

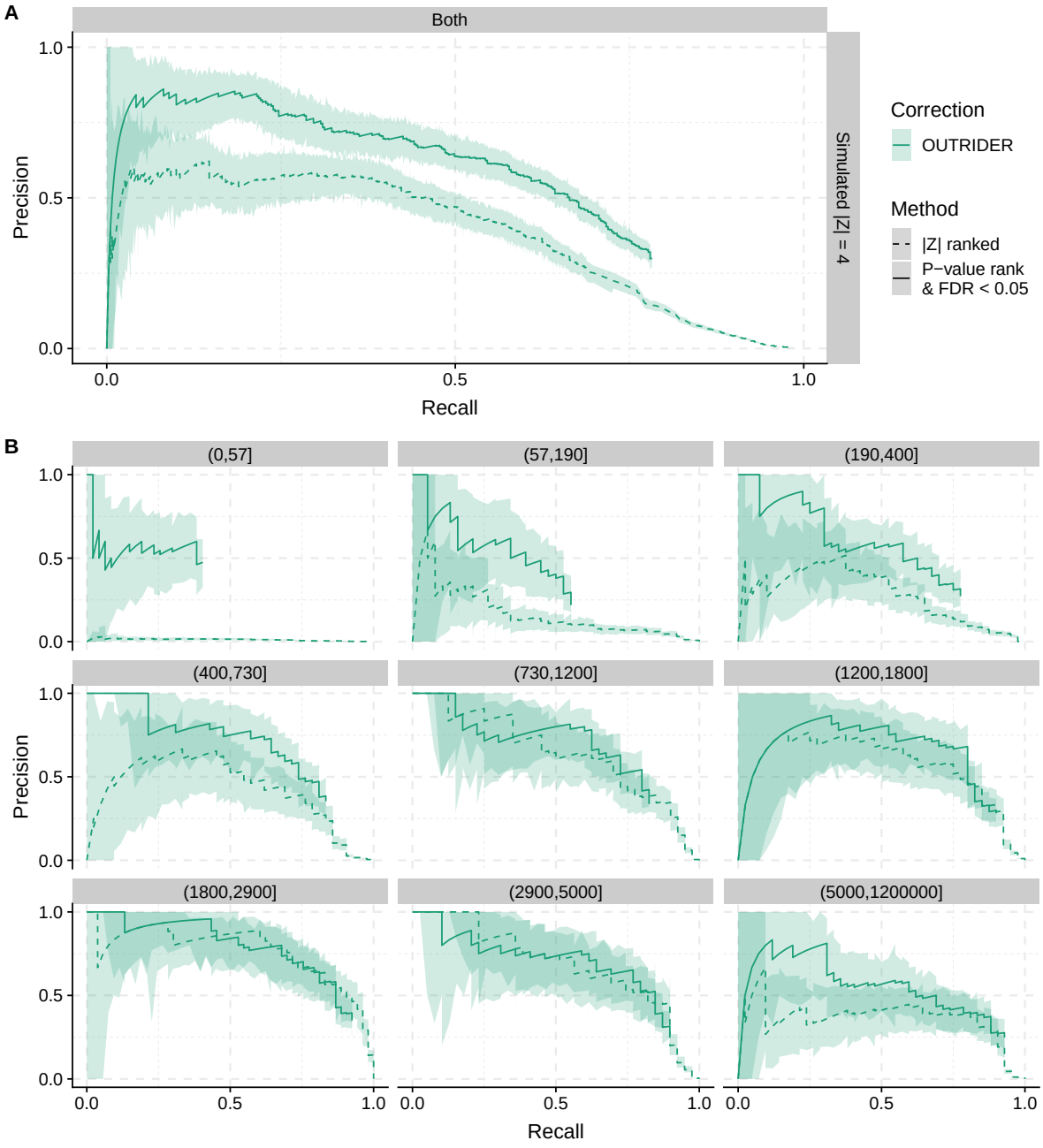

Figure A.8: Expression level dependent recall. Precision versus recall for artificially injected high and low expression outliers with a z score of 4 for OUTRIDER ranked by P values with FDR < 0.05 (solid) and ranked by z score (dashed). **(A)** For all the injected outliers. **(B)** Splitted into 9 bins, with equal number of read counts per bin, according to the mean expression level of the genes. Only a small fraction of the injected outliers was significant for the lowest bin, with a mean expression level smaller than 58. Adapted from Brechtmann et al. [2018].

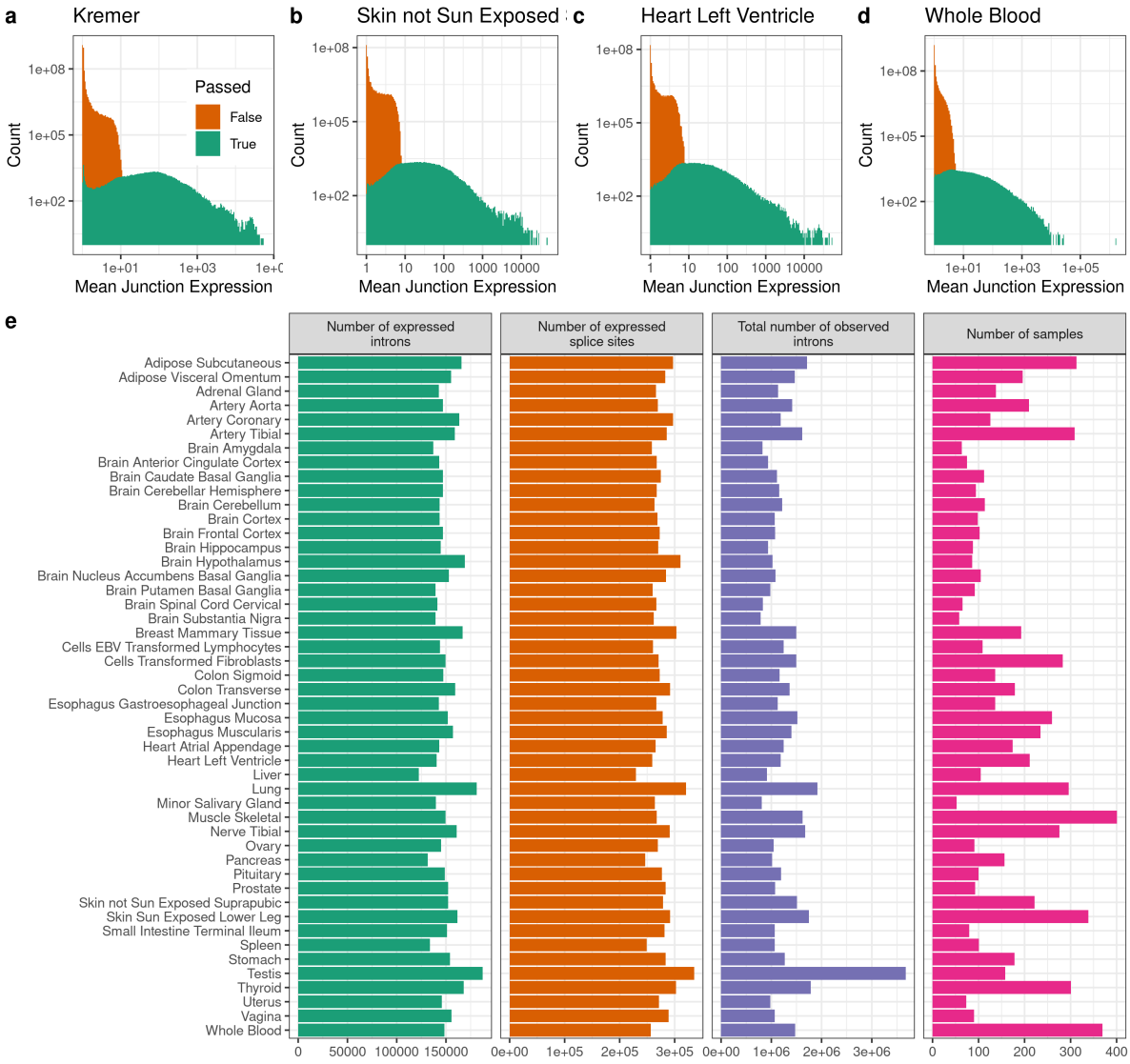

Figure A.9: Supplementary FiguresFigure S1: **(a)** Histogram of the raw intron coverage per sample-intron pair for the Kremer data set grouped according to the intron filter status. Green indicates that the intron passed the filter and orange indicates that the intron was filtered out. **(b)** Same as a but for the GTEx suprapubic skin tissue. **(c)** Same as a but for the GTEx left ventricle heart tissue. **(d)** Same as a but for the GTEx whole blood tissue. **(f)** Barplots of the number of introns passed the filtering, splice sites passed the filtering, observed introns, and samples per tissue within the GTEx dataset. Adapted from Mertes et al. [2021].

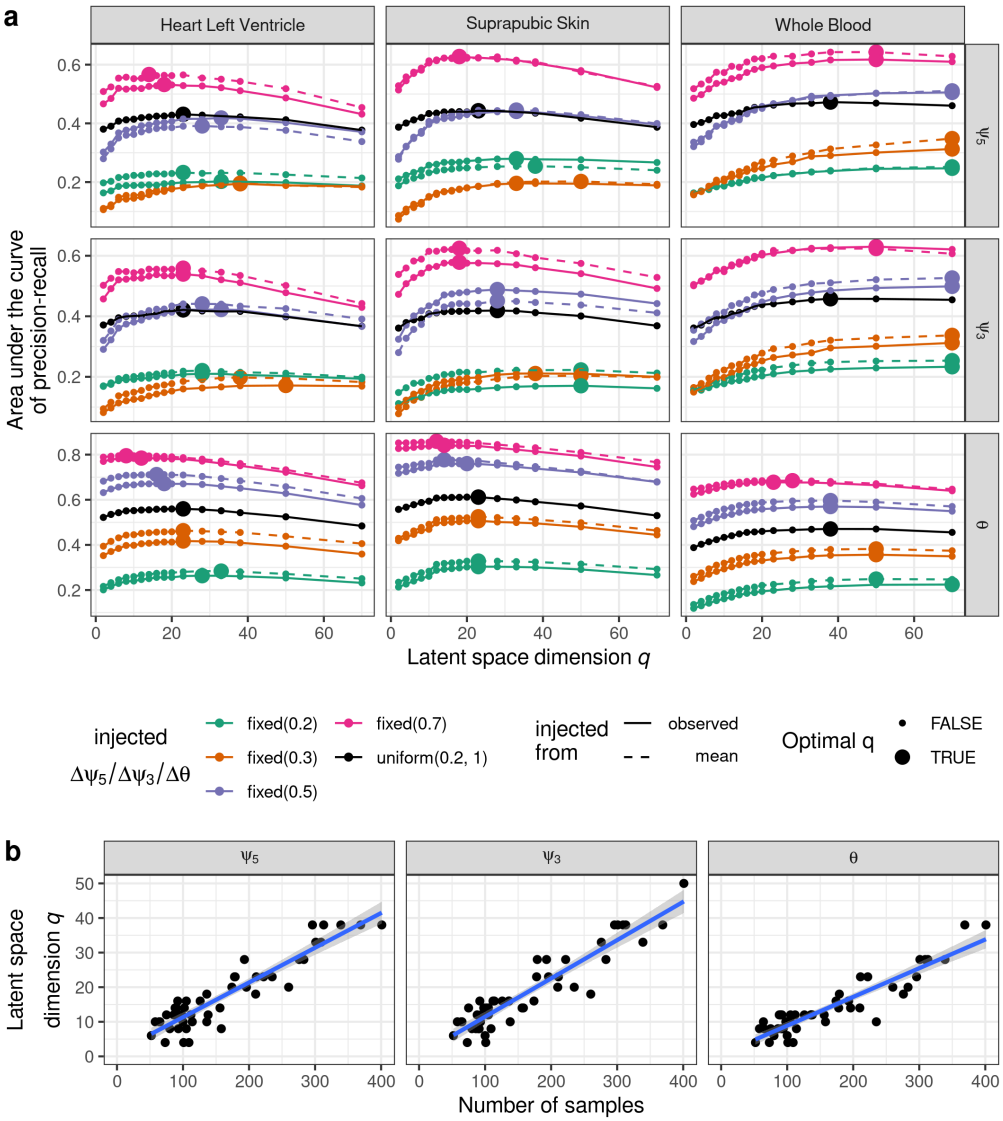

Figure A.10: Finding the optimal latent space dimension q. **(a)** Area under the precision-recall curve for recalling artificially injected outliers (y-axis) against latent space dimension q (x-axis) stratified by splicing metrics (rows) and three representative GTEx datasets (columns). Simulated outliers are generated using different scenarios: By shifting the splicing metrics away from its observed value (plain) or from its average across samples (dashed) and with shift of various amplitudes: 0.2 (green), 0.3 (orange), 0.5(purple) or 0.7 (pink) as well as with amplitudes drawn uniformly in [0.2,1] (black). For each scenario, the optimal latent space dimension q is marked with a thicker dot. **(b)** For each of the 48 GTEx tissues, the number of samples are plotted against the estimated latent space dimension. The data is stratified by the splicing metrics (columns). The blue line represents a linear regression fit and the gray band around it defines the 95% confidence interval of the fit. Adapted from Mertes et al. [2021].

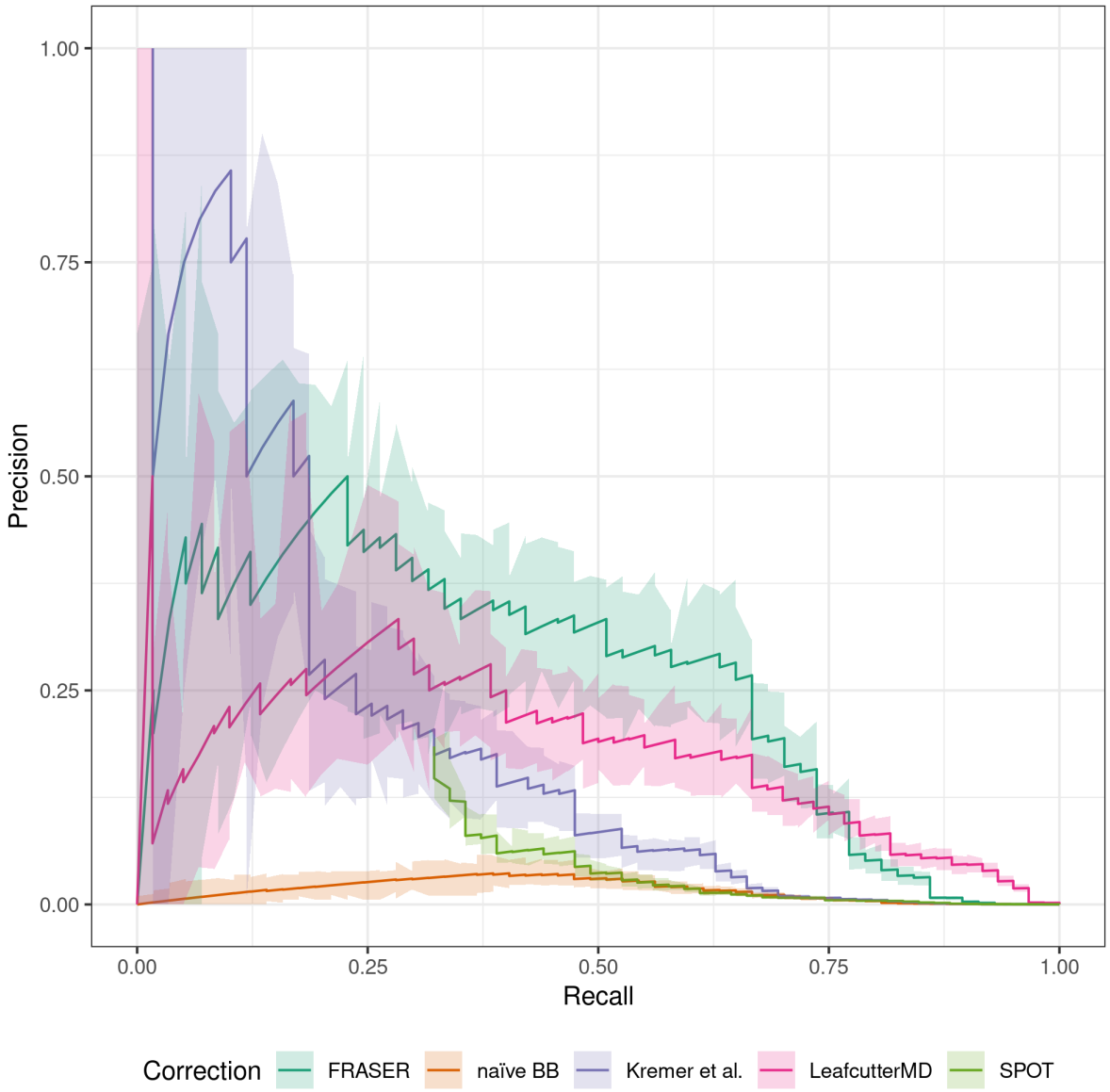

Figure A.11: Recall analysis of injected outliers by interchanging read counts of alternatively spliced genes between tissues. Proportion of simulated outliers among reported outliers (precision, y-axis) against the proportion of reported simulated outliers among all simulated outliers (recall, x-axis) for increasing BB P values computed using count ratio expectations based on FRASER (green) and on raw count ratios (orange, naïve BB) and Dirichlet-Multinomial P values computed using the LeafCutter adaptation (purple, Section 2.2.2) and the methods LeafcutterMD (pink) and SPOT (light green). The darker lines mark the precision-recall curves computed for the full dataset while the light ribbons around the curves depict 95% confidence bands estimated by bootstrapping. Abbreviations: BB, beta-binomial. Adapted from Mertes et al. [2021].

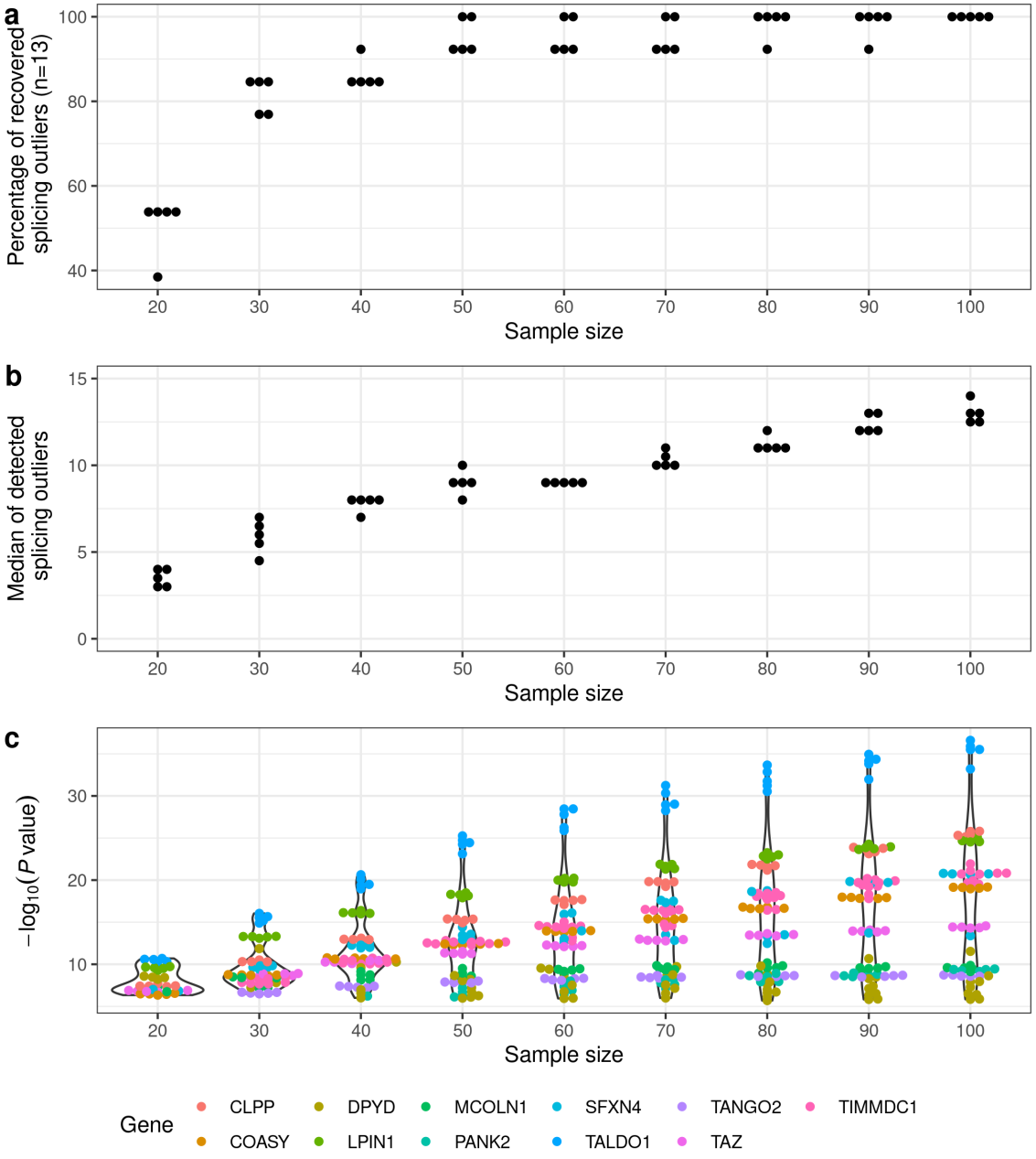

Figure A.12: Sample size analysis in the Kremer dataset. **(a)** The percentage of the recovered known disease-causing splicing outliers in the Kremer dataset (y-axis) is plotted against the used sample size (x-axis). The random sample selection was repeated 5 times (dots). **(b)** The median of splicing outliers across all samples (y-axis) is plotted against the used sample size (x-axis). **(c)** The negative $\log_{10}$ P value for all known disease-causing splicing outliers (y-axis) is plotted against the used sample size (x-axis). The color depicts the gene with a known splice defect. The violin depicts the density of the data points. Adapted from Mertes et al. [2021].

List of Figures

Acronyms

BAM	binary alignment map. 20, 53
BB	beta-binomial. 16, 28, 52–58, 61–63, 67, 80, 82
DNA	deoxyribonucleic acid. 2, 5–8
FDR	false discovery rate. 25, 39, 42, 43, 52, 56, 62
FPKM	fragments per kilobase per millions of reads. 35
FWER	family-wise error rate. 20, 22, 33, 56, 61, 64
GTEx	genotype-tissue expression. 28, 29, 34
INDEL	insertion or deletion of bases. 2, 6
MAE	mono-allelic expression. 16, 19, 25, 27, 67, 73, 74, 77, 78, 83
MAF	minor allele frequency. 23, 25, 26, 64, 103
NB	negative binomial. 25, 34, 35, 37, 39–43, 48, 54, 78, 80
NGS	next-generation sequencing. 2, 6, 8, 17, 77
OMIM	Online Mendelian Inheritance in Man. 4, 23, 27, 73
PCA	principal-component analysis. 38, 40–43, 45–48, 52, 55, 58, 62, 79, 80, 105, 106
PEER	probabilistic estimation of expression residuals. 34, 39–48, 79, 80, 105, 106
RIN	RNA integrity number. 34, 35, 52, 59
RNA	ribonucleic acid. 7, 8, 19, 21, 77
RNA-seq	RNA sequencing. 1, 4, 8, 10, 15, 16, 18–21, 25–28, 30, 33, 34, 36, 40, 41, 43, 48, 49, 51–53, 58, 66, 67, 71–73, 77–80, 82–84, 103
RPKM	reads per kilobase per million mapped reads. 34
SNV	single nucleotide variant. 2, 6, 19, 25–27, 103

VUS variant of unknown significance. 4, 17, 19, 23, 25–27, 30, 51, 72

WES whole-exome sequencing. 2–8, 16–19, 21, 25–27, 33, 71, 73, 77–79, 84, 103

WGS whole-genome sequencing. 2–8, 16–18, 21, 23, 24, 30, 33, 51, 73, 77, 79, 84

References

A. Abramowicz and M. Gos. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics*, 59(3):253–268, Aug. 2018. ISSN 1234-1983, 2190-3883. doi: 10.1007/s13353-018-0444-7. URL http://link.springer.com/10.1007/s13353-018-0444-7.

E. Adams and L. Frank. Metabolism of Proline and the Hydroxyprolines. *Annual Review of Biochemistry*, 49(1):1005–1061, June 1980. ISSN 0066-4154, 1545-4509. doi: 10.1146/annurev.bi.49.070180.005041. URL http://www.annualreviews.org/doi/10.1146/annurev.bi.49.070180.005041.

I. Adzhubei, D. M. Jordan, and S. R. Sunyaev. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, 0 7:Unit7.20, Jan. 2013. ISSN 1934-8266. doi: 10.1002/0471142905.hg0720s76. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4480630/.

J. K. Aicher, P. Jewell, J. Vaquero-Garcia, Y. Barash, and E. J. Bhoj. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genetics in Medicine*, Mar. 2020. ISSN 1098-3600, 1530-0366. doi: 10.1038/s41436-020-0780-y. URL http://www.nature.com/articles/s41436-020-0780-y.

S. Ait-El-Mkadem, M. Dayem-Quere, M. Gusic, A. Chaussenot, S. Bannwarth, B. François, E. Genin, K. Fragaki, C. Volker-Touw, C. Vasnier, V. Serre, K. van Gassen, F. Lespinasse, S. Richter, G. Eisenhofer, C. Rouzier, F. Mochel, A. De Saint-Martin, M.-T. Abi Warde, M. de Sain-van der Velde, J. Jans, J. Amiel, Z. Avsec, C. Mertes, T. Haack, T. Strom, T. Meitinger, P. Bonnen, R. Taylor, J. Gagneur, P. van Hasselt, A. Rötig, A. Delahodde, H. Prokisch, S. Fuchs, and V. Paquis-Flucklinger. Mutations in MDH2, Encoding a Krebs Cycle Enzyme, Cause Early-Onset Severe Encephalopathy. *American Journal of Human Genetics*, 100(1), 2017. ISSN 15376605. doi: 10.1016/j.ajhg.2016.11.014.

All of Us. The "All of Us" Research Program. *New England Journal of Medicine*, 381(7):668–676, Aug. 2019. ISSN 0028-4793. doi: 10.1056/NEJMsr1809937. URL https://doi.org/10.1056/NEJMsr1809937.

J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, 47(D1):

D1038–D1043, Jan. 2019. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gky1151. URL https://academic.oup.com/nar/article/47/D1/D1038/5184722.

S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, Oct. 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.133744.111. URL https://genome.cshlp.org/content/22/10/2008.

S. Anders, P. T. Pyl, and W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, Jan. 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu638. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu638.

B. Andrews, J. Carroll, S. Ding, I. M. Fearnley, and J. E. Walker. Assembly factors for the membrane arm of human complex I. *Proceedings of the National Academy of Sciences*, 110(47):18934–18939, Nov. 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1319247110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1319247110.

R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, June 2018. ISSN 1744-4292. doi: 10.15252/msb.20178124. URL http://www.embopress.org/doi/abs/10.15252/msb.20178124.

R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, and O. Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111, May 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02015-1. URL https://doi.org/10.1186/s13059-020-02015-1.

C. P. Austin, C. M. Cutillo, L. P. Lau, A. H. Jonker, A. Rath, D. Julkowska, D. Thomson, S. F. Terry, B. de Montleau, D. Ardigò, V. Hivert, K. M. Boycott, G. Baynam, P. Kaufmann, D. Taruscio, H. Lochmüller, M. Suematsu, C. Incerti, R. Draghia-Akli, I. Norstedt, L. Wang, H. J. Dawkins, and on behalf of the International Rare Diseases Research Consortium (IRDiRC). Future of Rare Diseases Research 2017-2027: An IRDiRC Perspective: Future of Rare Diseases Research 2017-2027. *Clinical and Translational Science*, 11(1):21–27, Jan. 2018. ISSN 17528054. doi: 10.1111/cts.12500. URL http://doi.wiley.com/10.1111/cts.12500.

M. B. Badsha, R. Li, B. Liu, Y. I. Li, M. Xian, N. E. Banovich, and A. Q. Fu. Imputation of single-cell gene expression with an autoencoder neural network. *Quantitative Biology*, 8(1):78–94, Mar. 2020. ISSN 2095-4697. doi: 10.1007/s40484-019-0192-7. URL https://doi.org/10.1007/s40484-019-0192-7.

M. N. Bainbridge, R. L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go, A. Delaney, M. Griffith, M. Hickenbotham, V. Magrini, E. R. Mardis, M. D. Sadar,

A. S. Siddiqui, M. A. Marra, and S. J. Jones. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7(1):246, Sept. 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-246. URL https://doi.org/10.1186/1471-2164-7-246.

M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, Nov. 2011. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3031. URL http://www.nature.com/articles/nrg3031.

V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, New York, 1974. ISBN 978-0-471-93094-5. URL https://www.wiley.com/en-us/Outliers+in+Statistical+Data%2C+3rd+Edition-p-9780471930945.

M. R. Baumgartner. Hyperammonemia with reduced ornithine, citrulline, arginine and proline: a new inborn error caused by a mutation in the gene encoding Delta1-pyrroline-5-carboxylate synthase. *Human Molecular Genetics*, 9(19):2853–2858, Nov. 2000. ISSN 14602083. doi: 10.1093/hmg/9.19.2853. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/9.19.2853.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289 – 300, 1995. ISSN 00359246. doi: 10.2307/2346101. URL http://www.jstor.org/stable/info/2346101.

Y. Benjamini and D. Yekutieli. The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*, 29(4):24, Apr. 2001. doi: 10.1214/aos/1013699998.

N. N. Borna, Y. Kishita, M. Kohda, S. C. Lim, M. Shimura, Y. Wu, K. Mogushi, Y. Yatsuka, H. Harashima, Y. Hisatomi, T. Fushimi, K. Ichimoto, K. Murayama, A. Ohtake, and Y. Okazaki. Mitochondrial ribosomal protein PTCD3 mutations cause oxidative phosphorylation defects with Leigh syndrome. *neurogenetics*, 20(1):9–25, Mar. 2019. ISSN 1364-6753. doi: 10.1007/s10048-018-0561-9. URL https://doi.org/10.1007/s10048-018-0561-9.

A. Bourdon, L. Minai, V. Serre, J.-P. Jais, E. Sarzi, S. Aubert, D. Chrétien, P. de Lonlay, V. Paquis-Flucklinger, H. Arakawa, Y. Nakamura, A. Munnich, and A. Rötig. Mutation of RRM2B , encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. *Nature Genetics*, 39(6):776–780, June 2007. ISSN 1546-1718. doi: 10.1038/ng2040. URL https://www.nature.com/articles/ng2040.

H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4):291–294, Sept. 1988. ISSN 1432-0770. doi: 10.1007/BF00332918. URL https://doi.org/10.1007/BF00332918.

K. M. Boycott and D. Ardigó. Addressing challenges in the diagnosis and treatment of rare genetic diseases. *Nature Reviews Drug Discovery*, 17(3):151–152, Mar. 2018. ISSN 1474-1784. doi: 10.1038/nrd.2017.246. URL `https://www.nature.com/articles/nrd.2017.246`.

K. M. Boycott, A. Rath, J. X. Chong, T. Hartley, F. S. Alkuraya, G. Baynam, A. J. Brookes, M. Brudno, A. Carracedo, J. T. den Dunnen, S. O. Dyke, X. Estivill, J. Goldblatt, C. Gonthier, S. C. Groft, I. Gut, A. Hamosh, P. Hieter, S. Höhn, M. E. Hurles, P. Kaufmann, B. M. Knoppers, J. P. Krischer, M. Macek, G. Matthijs, A. Olry, S. Parker, J. Paschall, A. A. Philippakis, H. L. Rehm, P. N. Robinson, P.-C. Sham, R. Stefanov, D. Taruscio, D. Unni, M. R. Vanstone, F. Zhang, H. Brunner, M. J. Bamshad, and H. Lochmüller. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics*, 100(5):695–705, May 2017. ISSN 00029297. doi: 10.1016/j.ajhg.2017.04.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929717301477`.

K. M. Boycott, T. Hartley, L. G. Biesecker, R. A. Gibbs, A. M. Innes, O. Riess, J. Belmont, S. L. Dunwoodie, N. Jojic, T. Lassmann, D. Mackay, I. K. Temple, A. Visel, and G. Baynam. A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. *Cell*, 177(1):32–37, Mar. 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.02.040. URL `https://www.sciencedirect.com/science/article/pii/S0092867419302235`.

U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24 (11):1774–1786, Nov. 2014. ISSN 1549-5469. doi: 10.1101/gr.177790.114.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, May 2016. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3519. URL `http://www.nature.com/articles/nbt.3519`.

F. Brechtmann, C. Mertes, A. Matusevičiūtė, V. A. Yépez, Ž. Avsec, M. Herzog, D. M. Bader, H. Prokisch, and J. Gagneur. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *The American Journal of Human Genetics*, 103(6):907–917, Dec. 2018. ISSN 00029297. doi: 10.1016/j.ajhg.2018.10.025. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929718304014`.

A. J. Brookes and P. N. Robinson. Human genotype–phenotype databases: aims, challenges and opportunities. *Nature Reviews Genetics*, 16(12):702–715, Dec. 2015. ISSN 1471-0064. doi: 10.1038/nrg3932. URL `https://www.nature.com/articles/nrg3932`.

C. A. Brownstein, I. A. Holm, R. Ramoni, and D. B. Goldstein. Data Sharing in the Undiagnosed Diseases Network. *Human Mutation*, 36(10):985–988, 2015. ISSN 1098-

1004. doi: https://doi.org/10.1002/humu.22840. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.22840.

C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, Apr. 1997. ISSN 00222836. doi: 10.1006/jmbi.1997.0951. URL https://linkinghub.elsevier.com/retrieve/pii/S0022283697909517.

R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, Sept. 1995. ISSN 1064-8275. doi: 10.1137/0916069. URL https://epubs.siam.org/doi/10.1137/0916069.

J. Casper, A. S. Zweig, C. Villarreal, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, D. Karolchik, A. S. Hinrichs, M. Haeussler, L. Guruvadoo, J. Navarro Gonzalez, D. Gibson, I. T. Fiddes, C. Eisenhart, M. Diekhans, H. Clawson, G. P. Barber, J. Armstrong, D. Haussler, R. M. Kuhn, and W. Kent. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, 46(D1):D762–D769, Jan. 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkx1020. URL https://academic.oup.com/nar/article/46/D1/D762/4584624.

J. C. Castle, C. Zhang, J. K. Shah, A. V. Kulkarni, A. Kalsotra, T. A. Cooper, and J. M. Johnson. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics*, 40(12):1416–1425, Dec. 2008. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.264. URL http://www.nature.com/articles/ng.264.

S. C. Chandrasekharappa, F. P. Lach, D. C. Kimble, A. Kamat, J. K. Teer, F. X. Donovan, E. Flynn, S. K. Sen, S. Thongthip, E. Sanborn, A. Smogorzewska, A. D. Auerbach, E. A. Ostrander, and and NISC Comparative Sequencing Program5. Massively parallel sequencing, aCGH, and RNA-Seq technologies provide a comprehensive molecular diagnosis of Fanconi anemia. *Blood*, 121(22):e138–e148, May 2013. ISSN 0006-4971. doi: 10.1182/blood-2012-12-474585. URL https://doi.org/10.1182/blood-2012-12-474585.

W. Chauvenet. *A manual of spherical and practical astronomy.* Philadelphia, J. B. Lippincott & co.; London, Trübner & co., 1863. URL http://archive.org/details/amanualspherica06chaugoog.

J. Cheng, T. Y. D. Nguyen, K. J. Cygan, M. H. Çelik, W. G. Fairbrother, Ž. Avsec, and J. Gagneur. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biology*, 20(1):48, Dec. 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1653-z. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1653-z.

J. Cheng, M. H. Çelik, A. Kundaje, and J. Gagneur. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biology*, 22(1):94, Mar. 2021.

ISSN 1474-760X. doi: 10.1186/s13059-021-02273-7. URL https://doi.org/10.1186/s13059-021-02273-7.

F. Cheung, B. J. Haas, S. M. Goldberg, G. D. May, Y. Xiao, and C. D. Town. Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, 7(1):272, Oct. 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-272. URL https://doi.org/10.1186/1471-2164-7-272.

J. X. Chong, K. J. Buckingham, S. N. Jhangiani, C. Boehm, N. Sobreira, J. D. Smith, T. M. Harrell, M. J. McMillin, W. Wiszniewski, T. Gambin, Z. H. Coban Akdemir, K. Doheny, A. F. Scott, D. Avramopoulos, A. Chakravarti, J. Hoover-Fong, D. Mathews, P. D. Witmer, H. Ling, K. Hetrick, L. Watkins, K. E. Patterson, F. Reinier, E. Blue, D. Muzny, M. Kircher, K. Bilguvar, F. López-Giráldez, V. R. Sutton, H. K. Tabor, S. M. Leal, M. Gunel, S. Mane, R. A. Gibbs, E. Boerwinkle, A. Hamosh, J. Shendure, J. R. Lupski, R. P. Lifton, D. Valle, D. A. Nickerson, and M. J. Bamshad. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics*, 97(2):199–215, Aug. 2015. ISSN 00029297. doi: 10.1016/j.ajhg.2015.06.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929715002451.

M. M. Clark, Z. Stark, L. Farnaes, T. Y. Tan, S. M. White, D. Dimmock, and S. F. Kingsmore. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Medicine*, 3(1):16, Dec. 2018. ISSN 2056-7944. doi: 10.1038/s41525-018-0053-8. URL http://www.nature.com/articles/s41525-018-0053-8.

D. Collett and T. Lewis. The Subjective Nature of Outlier Rejection Procedures. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(3):228–237, 1976. ISSN 0035-9254. doi: 10.2307/2347230. URL https://www.jstor.org/stable/2347230.

R. D. Cook. Detection of Influential Observation in Linear Regression. *Technometrics*, 19 (1):15–18, 1977. ISSN 0040-1706. doi: 10.2307/1268249. URL https://www.jstor.org/stable/1268249.

D. Cousineau and S. Chartier. Outliers detection and treatment: a review. *International Journal of Psychological Research*, 3(1):58–67, June 2010. ISSN 2011-7922. doi: 10.21500/20112084.844. URL https://revistas.usb.edu.co/index.php/IJPR/article/view/844.

M. Coutelier, C. Goizet, A. Durr, F. Habarou, S. Morais, A. Dionne-Laporte, F. Tao, J. Konop, M. Stoll, P. Charles, M. Jacoupy, R. Matusiak, I. Alonso, C. Tallaksen, M. Mairey, M. Kennerson, M. Gaussen, R. Schule, M. Janin, F. Morice-Picard, C. M. Durand, C. Depienne, P. Calvas, P. Coutinho, J.-M. Saudubray, G. Rouleau, A. Brice, G. Nicholson, F. Darios, J. L. Loureiro, S. Zuchner, C. Ottolenghi, F. Mochel, and G. Stevanin. Alteration of ornithine metabolism leads to dominant and recessive

hereditary spastic paraplegia. *Brain*, 138(8):2191–2205, Aug. 2015. ISSN 0006-8950, 1460-2156. doi: 10.1093/brain/awv143. URL `https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awv143`.

B. B. Cummings, J. L. Marshall, T. Tukiainen, M. Lek, S. Donkervoort, A. R. Foley, V. Bolduc, L. B. Waddell, S. A. Sandaradura, G. L. O'Grady, E. Estrella, H. M. Reddy, F. Zhao, B. Weisburd, K. J. Karczewski, A. H. O'Donnell-Luria, D. Birnbaum, A. Sarkozy, Y. Hu, H. Gonorazky, K. Claeys, H. Joshi, A. Bournazos, E. C. Oates, R. Ghaoui, M. R. Davis, N. G. Laing, A. Topf, P. B. Kang, A. H. Beggs, K. N. North, V. Straub, J. J. Dowling, F. Muntoni, N. F. Clarke, S. T. Cooper, C. G. Bönnemann, and D. G. MacArthur. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine*, 9(386):eaal5209, Apr. 2017. ISSN 1946-6234. doi: 10.1126/scitranslmed.aal5209.

B. B. Cummings, K. J. Karczewski, J. A. Kosmicki, E. G. Seaby, N. A. Watts, M. Singer-Berk, J. M. Mudge, J. Karjalainen, F. K. Satterstrom, A. H. O'Donnell-Luria, T. Poterba, C. Seed, M. Solomonson, J. Alföldi, G. A. D. P. Team, G. A. D. Consortium, M. J. Daly, and D. G. MacArthur. Transcript expression-aware annotation improves rare variant interpretation. *Nature*, 581(7809):452–458, May 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2329-2. URL `http://www.nature.com/articles/s41586-020-2329-2`.

Y. Dai and G. Wang. Analyzing Tongue Images Using a Conceptual Alignment Deep Autoencoder. *IEEE Access*, 6:5962–5972, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2017.2788849.

R. B. Dean and W. J. Dixon. Simplified Statistics for Small Numbers of Observations. *Analytical Chemistry*, 23(4):636–638, Apr. 1951. ISSN 0003-2700. doi: 10.1021/ac60052a025. URL `https://doi.org/10.1021/ac60052a025`.

D. S. DeLuca, J. Z. Levin, A. Sivachenko, T. Fennell, M.-D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, June 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts196. URL `https://doi.org/10.1093/bioinformatics/bts196`.

F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research*, 37(9):e67–e67, May 2009. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gkp215. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp215`.

M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and on behalf of The

French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, Nov. 2013. ISSN 1467-5463. doi: 10.1093/bib/bbs046. URL https://doi.org/10.1093/bib/bbs046.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. ISSN 13674803. doi: 10.1093/bioinformatics/bts635.

R. I. Dogan, L. Getoor, W. J. Wilbur, and S. M. Mount. SplicePort–An interactive splice-site analysis tool. *Nucleic Acids Research*, 35(Web Server):W285–W291, May 2007. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkm407. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm407.

H. L. Drexler, K. Choquet, and L. S. Churchman. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Molecular Cell*, 77 (5):985–998.e8, Mar. 2020. ISSN 1097-2765. doi: 10.1016/j.molcel.2019.11.017. URL http://www.sciencedirect.com/science/article/pii/S1097276519308652.

J. M. Ellingford, H. B. Thomas, C. Rowlands, G. Arno, G. Beaman, B. Gomes-Silva, C. Campbell, N. Gossan, C. Hardcastle, K. Webb, C. O'Callaghan, R. A. Hirst, S. Ramsden, E. Jones, J. Clayton-Smith, A. R. Webster, Genomics England Research Consortium, R. T. O'Keefe, W. G. Newman, and G. C. Black. Functional and in-silico interrogation of rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *bioRxiv*, page 781088, Sept. 2019. doi: 10.1101/781088. URL https://www.biorxiv.org/content/10.1101/781088v1.

S. J. Emrich, W. B. Barbazuk, L. Li, and P. S. Schnable. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, 17(1):69–73, Jan. 2007. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.5145806. URL https://genome.cshlp.org/content/17/1/69.

L. Eng, G. Coutinho, S. Nahas, G. Yeo, R. Tanouye, M. Babaei, T. Dörk, C. Burge, and R. A. Gatti. Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: Maximum entropy estimates of splice junction strengths. *Human Mutation*, 23(1):67–76, 2004. ISSN 1098-1004. doi: https://doi.org/10.1002/humu.10295. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.10295.

G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390, Jan. 2019. ISSN 2041-1723. doi: 10.1038/s41467-018-07931-2. URL https://www.nature.com/articles/s41467-018-07931-2.

N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K.

Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan, GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, and A. Battle. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science*, 369(6509):eaaz5900, Sept. 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaz5900. URL `https://www.sciencemag.org/lookup/doi/10.1126/science.aaz5900`.

L. Ferri, C. Dionisi-Vici, R. Taurisano, F. M. Vaz, R. Guerrini, and A. Morrone. When silence is noise: infantile-onset Barth syndrome caused by a synonymous substitution affecting TAZ gene transcription. *Clinical Genetics*, 90(5):461–465, 2016. ISSN 1399-0004. doi: 10.1111/cge.12756.

P. Filzmoser and M. Gregorich. Multivariate Outlier Detection in Applied Data Analysis: Global, Local, Compositional and Cellwise Outliers. *Mathematical Geosciences*, 52 (8):1049–1066, Nov. 2020. ISSN 1874-8953. doi: 10.1007/s11004-020-09861-6. URL `https://doi.org/10.1007/s11004-020-09861-6`.

B. Fischer-Zirnsak, N. Escande-Beillard, J. Ganesh, Y. Tan, M. Al Bughaili, A. Lin, I. Sahai, P. Bahena, S. Reichert, A. Loh, G. Wright, J. Liu, E. Rahikkala, E. Pivnick, A. Choudhri, U. Krüger, T. Zemojtel, C. van Ravenswaaij Arts, R. Mostafavi, I. Stolte-Dijkstra, S. Symoens, L. Pajunen, L. Al-Gazali, D. Meierhofer, P. Robinson, S. Mundlos, C. Villarroel, P. Byers, A. Masri, S. Robertson, U. Schwarze, B. Callewaert, B. Reversade, and U. Kornak. Recurrent De Novo Mutations Affecting Residue Arg138 of Pyrroline-5-Carboxylate Synthase Cause a Progeroid Form of Autosomal-Dominant Cutis Laxa. *The American Journal of Human Genetics*, 97(3):483–492, Sept. 2015. ISSN 00029297. doi: 10.1016/j.ajhg.2015.08.001. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929715003225`.

R. A. Fisher. *Statistical Methods for Research Workers, 14th edition.* Oliver & Boyd, 14th edition, 1970. ISBN 978-0-05-002170-5.

I. F. A. C. Fokkema, P. E. M. Taschner, G. C. P. Schaafsma, J. Celli, J. F. J. Laros, and J. T. d. Dunnen. LOVD v.2.0: the next generation in gene variant databases. *Human Mutation*, 32(5):557–563, 2011. ISSN 1098-1004. doi: https://doi.org/10.1002/humu.21438. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.21438`.

L. Frésard, C. Smail, N. M. Ferraro, N. A. Teran, X. Li, K. S. Smith, D. Bonner, K. D. Kernohan, S. Marwaha, Z. Zappala, B. Balliu, J. R. Davis, B. Liu, C. J. Prybol, J. N. Kohler, D. B. Zastrow, C. M. Reuter, D. G. Fisk, M. E. Grove, J. M. Davidson, T. Hartley, R. Joshi, B. J. Strober, S. Utiramerur, L. Lind, E. Ingelsson, A. Battle, G. Bejerano, J. A. Bernstein, E. A. Ashley, K. M. Boycott, J. D. Merker, M. T. Wheeler, and S. B. Montgomery. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*, 25(6):911–919, June 2019. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-019-0457-8. URL `http://www.nature.com/articles/s41591-019-0457-8`.

W. A. Gahl, A. L. Wise, and E. A. Ashley. The Undiagnosed Diseases Network of the National Institutes of Health: A National Extension. *JAMA*, 314(17):1797, Nov. 2015. ISSN 0098-7484. doi: 10.1001/jama.2015.12249. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.12249.

R. C. Gallo. Molecular Biology: Reverse Transcriptase, the DNA Polymerase of Oncogenic RNA Viruses. *Nature*, 234(5326):194–198, Nov. 1971. ISSN 1476-4687. doi: 10.1038/234194a0. URL https://www.nature.com/articles/234194a0.

Genomics England. The National Genomics Research and Healthcare Knowledgebase. *figshare*, Dec. 2017. doi: 10.6084/m9.figshare.4530893.v5. URL https://figshare.com/articles/dataset/GenomicEnglandProtocol_pdf/4530893/5.

C. Gilissen, A. Hoischen, H. G. Brunner, and J. A. Veltman. Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, 20(5):490–497, May 2012. ISSN 1018-4813, 1476-5438. doi: 10.1038/ejhg.2011.258. URL http://www.nature.com/articles/ejhg2011258.

H. D. Gonorazky, S. Naumenko, A. K. Ramani, V. Nelakuditi, P. Mashouri, P. Wang, D. Kao, K. Ohri, S. Viththiyapaskaran, M. A. Tarnopolsky, K. D. Mathews, S. A. Moore, A. N. Osorio, D. Villanova, D. U. Kemaladewi, R. D. Cohn, M. Brudno, and J. J. Dowling. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *The American Journal of Human Genetics*, 104(3):466–483, Mar. 2019. ISSN 00029297. doi: 10.1016/j.ajhg.2019.01.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929719300126.

G. S. Gorman, P. F. Chinnery, S. DiMauro, M. Hirano, Y. Koga, R. McFarland, A. Suomalainen, D. R. Thorburn, M. Zeviani, and D. M. Turnbull. Mitochondrial diseases. *Nature Reviews Disease Primers*, 2(1):16080, Dec. 2016. ISSN 2056-676X. doi: 10.1038/nrdp.2016.80. URL http://www.nature.com/articles/nrdp201680.

D. T. Grozdic and S. T. Jovicic. Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(12):2313–2322, Dec. 2017. ISSN 2329-9290. doi: 10.1109/TASLP.2017.2738559. URL https://doi.org/10.1109/TASLP.2017.2738559.

F. E. Grubbs. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1):1–21, Feb. 1969. ISSN 0040-1706. doi: 10.1080/00401706.1969.10490657. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657.

B. Grüning, R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, and J. Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, July 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0046-7. URL https://www.nature.com/articles/s41592-018-0046-7.

GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, Oct. 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature24277. URL http://www.nature.com/articles/nature24277.

V. Guarani, J. Paulo, B. Zhai, E. L. Huttlin, S. P. Gygi, and J. W. Harper. TIM-MDC1/C3orf1 Functions as a Membrane-Embedded Mitochondrial Complex I Assembly Factor through Association with the MCIA Complex. *Molecular and Cellular Biology*, 34(5):847–861, Mar. 2014. ISSN 0270-7306, 1098-5549. doi: 10.1128/MCB.01551-13. URL https://mcb.asm.org/content/34/5/847.

T. B. Haack, K. Danhauser, B. Haberberger, J. Hoser, V. Strecker, D. Boehm, G. Uziel, E. Lamantea, F. Invernizzi, J. Poulton, B. Rolinski, A. Iuso, S. Biskup, T. Schmidt, H.-W. Mewes, I. Wittig, T. Meitinger, M. Zeviani, and H. Prokisch. Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nature Genetics*, 42 (12):1131–1134, Dec. 2010. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.706. URL http://www.nature.com/articles/ng.706.

I. S. Hagemann. Overview of Technical Aspects and Chemistries of Next-Generation Sequencing. In *Clinical Genomics*, pages 3–19. Elsevier, 2015. ISBN 978-0-12-404748-8. doi: 10.1016/B978-0-12-404748-8.00001-0. URL https://linkinghub.elsevier.com/retrieve/pii/B9780124047488000010.

B. Hagl, B. D. Spielberger, S. Thoene, S. Bonnal, C. Mertes, C. Winter, I. J. Nijman, S. Verduin, A. C. Eberherr, A. Puel, D. Schindler, J. Ruland, T. Meitinger, J. Gagneur, J. S. Orange, M. E. van Gijn, and E. D. Renner. Somatic alterations compromised molecular diagnosis of DOCK8 hyper-IgE syndrome caused by a novel intronic splice site mutation. *Scientific Reports*, 8(1):16719, Dec. 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-34953-z. URL http://www.nature.com/articles/s41598-018-34953-z.

T. Halperin, B. Zheng, H. Itzhaki, A. K. Clarke, and Z. Adam. Plant mitochondria contain proteolytic and regulatory subunits of the ATP-dependent Clp protease. *Plant Molecular Biology*, 45(4):461–468, 2001. ISSN 01674412. doi: 10.1023/A:1010677220323. URL http://link.springer.com/10.1023/A:1010677220323.

K. D. Hansen, R. A. Irizarry, and Z. Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, 2012. doi: 10.1093/biostatistics/kxr054.

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, Sept. 2012. ISSN 1088-9051. doi: 10.1101/gr.135350.111.

G. E. Hinton and R. S. Zemel. Autoencoders, Minimum Description Length and Helmholtz Free Energy. In J. D. Cowan, G. Tesauro, and Alspector J., editors, *Advances in Neural Information Processing Systems 6*, volume Morgan-Kau, pages 3–10. Morgan-Kaufmann, 1994.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. ISSN 00063444. doi: 10.1093/biomet/75.4.800.

V. J. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, Oct. 2004. ISSN 1573-7462. doi: 10.1007/s10462-004-4304-y. URL `https://doi.org/10.1007/s10462-004-4304-y`.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL `http://www.jstor.org/stable/4615733`.

E. Holzerova, K. Danhauser, T. B. Haack, L. S. Kremer, M. Melcher, I. Ingold, S. Kobayashi, C. Terrile, P. Wolf, J. Schaper, E. Mayatepek, F. Baertling, J. P. Friedmann Angeli, M. Conrad, T. M. Strom, T. Meitinger, H. Prokisch, and F. Distelmaier. Human thioredoxin 2 deficiency impairs mitochondrial redox homeostasis and causes early-onset neurodegeneration. *Brain*, 139(2):346–354, Feb. 2016. ISSN 1460-2156, 0006-8950. doi: 10.1093/brain/awv350. URL `https://academic.oup.com/brain/article/139/2/346/1753955`.

H. Hotelling. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, Aug. 1931. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177732979. URL `https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-2/issue-3/The-Generalization-of-Students-Ratio/10.1214/aoms/1177732979.full`.

R. H. Houtkooper, M. Turkenburg, B. T. Poll-The, D. Karall, C. Pérez-Cerdá, A. Morrone, S. Malvagia, R. J. Wanders, W. Kulik, and F. M. Vaz. The enigmatic role of tafazzin in cardiolipin metabolism. *Biochimica Et Biophysica Acta*, 1788(10):2003–2014, Oct. 2009. ISSN 0006-3002. doi: 10.1016/j.bbamem.2009.07.009.

F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler. The UCSC Known Genes. *Bioinformatics*, 22(9):1036–1046, May 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl048.

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, Mar. 1964. ISSN 0003-4851. doi: 10.1214/aoms/1177703732. URL `http://projecteuclid.org/euclid.aoms/1177703732`.

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A.

Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, Feb. 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3252.

K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K.-H. Farh. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176 (3):535–548.e24, Jan. 2019. ISSN 00928674. doi: 10.1016/j.cell.2018.12.015. URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867418316295`.

E. Jenkinson, A. Rehman, T. Walsh, J. Clayton-Smith, K. Lee, R. Morell, M. Drummond, S. Khan, M. Naeem, B. Rauf, N. Billington, J. Schultz, J. Urquhart, M. Lee, A. Berry, N. Hanley, S. Mehta, D. Cilliers, P. Clayton, H. Kingston, M. Smith, T. Warner, G. Black, D. Trump, J. Davis, W. Ahmad, S. Leal, S. Riazuddin, M.-C. King, T. Friedman, and W. Newman. Perrault Syndrome Is Caused by Recessive Mutations in CLPP, Encoding a Mitochondrial ATP-Dependent Chambered Protease. *The American Journal of Human Genetics*, 92(4):605–613, Apr. 2013. ISSN 00029297. doi: 10.1016/j.ajhg.2013.02.013. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929713001080`.

E. M. Jenkinson, J. Clayton-Smith, S. Mehta, C. Bennett, W. Reardon, A. Green, S. H. S. Pearce, G. Michele, G. S. Conway, D. Cilliers, N. Moreton, J. R. E. Davis, D. Trump, and W. G. Newman. Perrault syndrome: further evidence for genetic heterogeneity. *Journal of Neurology*, 259(5):974–976, May 2012. ISSN 0340-5354, 1432-1459. doi: 10.1007/s00415-011-6285-5. URL `http://link.springer.com/10.1007/s00415-011-6285-5`.

G. Jenkinson, Y. I. Li, S. Basu, M. A. Cousin, G. R. Oliver, and E. W. Klee. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics*, pages 1–7, Apr. 2020. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btaa259. URL `https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa259/5823301`.

X. Jian, E. Boerwinkle, and X. Liu. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, 42(22):13534–13544, Dec. 2014. ISSN 0305-1048. doi: 10.1093/nar/gku1206. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku1206`.

Y. Kapustin, E. Chan, R. Sarkar, F. Wong, I. Vorechovsky, R. M. Winston, T. Tatusova, and N. J. Dibb. Cryptic splice sites and split genes. *Nucleic acids research*, 39 (14):5837–44, Aug. 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr203. URL `http://www.ncbi.nlm.nih.gov/pubmed/21470962`.

K. J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D. M. Ruderfer, D. Kavanagh, T. Hamamsy, M. Lek, K. E. Samocha, B. B. Cummings, D. Birnbaum, The Exome Aggregation Consortium, M. J. Daly, and D. G. MacArthur. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research*, 45(D1):D840–D845, Jan. 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw971. URL `https://doi.org/10.1093/nar/gkw971`.

K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, G. A. D. Consortium, B. M. Neale, M. J. Daly, and D. G. MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809): 434–443, May 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2308-7. URL `http://www.nature.com/articles/s41586-020-2308-7`.

Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12): 1009–1015, Dec. 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1528. URL `http://www.nature.com/articles/nmeth.1528`.

Y. Katz, E. T. Wang, J. Silterra, S. Schwartz, B. Wong, H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, E. M. Airoldi, and C. B. Burge. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, 31(14): 2400–2402, July 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btv034. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv034`.

N. Kenmochi, T. Suzuki, T. Uechi, M. Magoori, M. Kuniba, S. Higa, K. Watanabe, and T. Tanaka. The Human Mitochondrial Ribosomal Protein Genes: Mapping of 54 Genes to the Chromosomes and Implications for Human Disorders. *Genomics*, 77 (1):65–70, Sept. 2001. ISSN 0888-7543. doi: 10.1006/geno.2001.6622. URL `https://www.sciencedirect.com/science/article/pii/S0888754301966224`.

K. D. Kernohan, L. Frésard, Z. Zappala, T. Hartley, K. S. Smith, J. Wagner, H. Xu, A. McBride, P. R. Bourque, C. C. Consortium, S. A. L. Bennett, D. A. Dyment, K. M. Boycott, S. B. Montgomery, and J. Warman Chardon. Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy. *Human Mutation*, 38(6):611–614, June 2017. ISSN 10597794. doi: 10.1002/humu.23211. URL `http://doi.wiley.com/10.1002/humu.23211`.

Y. L. Khodor, J. Rodriguez, K. C. Abruzzi, C.-H. A. Tang, M. T. Marr, and M. Rosbash. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development*, 25(23):2502–2512, Dec. 2011. ISSN 1549-5477. doi: 10.1101/gad.178962.111.

S. Kinalis, F. C. Nielsen, O. Winther, and F. O. Bagger. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. *BMC Bioinformatics*, 20(1):379, July 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2952-9. URL `https://doi.org/10.1186/s12859-019-2952-9`.

M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–5, Mar. 2014. ISSN 1546-1718. doi: 10.1038/ng.2892. URL `http://www.ncbi.nlm.nih.gov/pubmed/24487276`.

R. Kopajtich, D. Smirnov, S. L. Stenton, S. Loipfinger, C. Meng, I. F. Scheller, P. Freisinger, R. Baski, R. Berutti, J. Behr, M. Bucher, F. Distelmaier, M. Gusic, M. Hempel, L. Kulterer, J. Mayr, T. Meitinger, C. Mertes, M. D. Metodiev, A. Nadel, A. Nasca, A. Ohtake, Y. Okazaki, R. Olsen, D. Piekutowska-Abramczuk, A. Rötig, R. Santer, D. Schindler, A. Slama, C. Staufner, T. Strom, P. Verloo, J.-C. von Kleist-Retzow, S. B. Wortmann, V. A. Yépez, C. Lamperti, D. Ghezzi, K. Murayama, C. Ludwig, J. Gagneur, and H. Prokisch. Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. *medRxiv*, Mar. 2021. doi: 10.1101/2021.03.09.21253187. URL `http://medrxiv.org/lookup/doi/10.1101/2021.03.09.21253187`.

L. S. Kremer, D. M. Bader, C. Mertes, R. Kopajtich, G. Pichler, A. Iuso, T. B. Haack, E. Graf, T. Schwarzmayr, C. Terrile, E. Koňaříková, B. Repp, G. Kastenmüller, J. Adamski, P. Lichtner, C. Leonhardt, B. Funalot, A. Donati, V. Tiranti, A. Lombes, C. Jardel, D. Gläser, R. W. Taylor, D. Ghezzi, J. A. Mayr, A. Rötig, P. Freisinger, F. Distelmaier, T. M. Strom, T. Meitinger, J. Gagneur, and H. Prokisch. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications*, 8:15824, June 2017. ISSN 2041-1723. doi: 10.1038/ncomms15824. URL `http://dx.doi.org/10.1038/ncomms15824`.

P. Kumar, S. Henikoff, and P. C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, 4(7):1073–81, Jan. 2009. ISSN 1750-2799. doi: 10.1038/nprot.2009.86. URL `http://dx.doi.org/10.1038/nprot.2009.86`.

N. J. Lake, B. D. Webb, D. A. Stroud, T. R. Richman, B. Ruzzenente, A. G. Compton, H. S. Mountford, J. Pulman, C. Zangarelli, M. Rio, N. Boddaert, Z. Assouline, M. D. Sherpa, E. E. Schadt, S. M. Houten, J. Byrnes, E. M. McCormick, Z. Zolkipli-Cunningham, K. Haude, Z. Zhang, K. Retterer, R. Bai, S. E. Calvo, V. K. Mootha, J. Christodoulou, A. Rötig, A. Filipovska, I. Cristian, M. J. Falk, M. D. Metodiev, and

D. R. Thorburn. Biallelic Mutations in MRPS34 Lead to Instability of the Small Mitoribosomal Subunit and Leigh Syndrome. *The American Journal of Human Genetics*, 101(2):239–254, Aug. 2017. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2017.07.005. URL `https://www.cell.com/ajhg/abstract/S0002-9297(17)30283-5`.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, International Human Genome Sequencing Consortium, C. f. G. R.

Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, I. o. M. B. Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, T. I. f. S. B. Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, L. A. H. G. C. Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, a. i. i. l. u. o. h. *Genome Analysis Group (listed in alphabetical order, U. N. I. o. H. Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, K. U. S. o. M. Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, U. D. o. E. Office of Science, and The Wellcome Trust:. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001. ISSN 1476-4687. doi: 10.1038/35057062. URL `https://www.nature.com/articles/35057062`.

M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, and D. R. Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, Jan. 2018. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkx1153. URL `http://academic.oup.com/nar/article/46/D1/D1062/4641904`.

T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, The Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468): 506–511, Sept. 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12531. URL `http://www.nature.com/articles/nature12531`.

M. Lawrence, W. Huber, H. Pag??s, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):e1003118, Aug. 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003118.

Y. Lecun. Modeles connexionnistes de l'apprentissage (connectionist learning models). *Université Pierre et Marie Curie, Paris*, page PhD thesis, 1987.

H. Lee, A. Y. Huang, L.-k. Wang, A. J. Yoon, G. Renteria, A. Eskin, R. H. Signer, N. Dorrani, S. Nieves-Rodriguez, J. Wan, E. D. Douine, J. D. Woods, E. C. Dell'Angelica, B. L. Fogel, M. G. Martin, M. J. Butte, N. H. Parker, R. T. Wang, P. B. Shieh, D. A. Wong, N. Gallant, K. E. Singh, Y. J. Tavyev Asher, J. S. Sinsheimer, D. Krakow, S. K. Loo, P. Allard, J. C. Papp, Undiagnosed Diseases Network, C. G. S. Palmer, J. A. Martinez-Agosto, and S. F. Nelson. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. *Genetics in Medicine*, Oct. 2019. ISSN 1098-3600, 1530-0366. doi: 10.1038/s41436-019-0672-1. URL `http://www.nature.com/articles/s41436-019-0672-1`.

K. K. Lee, M. Shimoji, Q. S. Hossain, H. Sunakawa, and Y. Aniya. Novel function of glutathione transferase in rat liver mitochondrial membrane: Role for cytochrome c release from mitochondria. *Toxicology and Applied Pharmacology*, 232 (1):109–118, Oct. 2008. ISSN 0041008X. doi: 10.1016/j.taap.2008.06.005. URL `https://linkinghub.elsevier.com/retrieve/pii/S0041008X08002548`.

J. T. Leek and J. D. Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics*, 3(9):e161, Sept. 2007. ISSN 1553-7404. doi: 10.1371/journal.pgen.0030161. URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030161`.

M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, and D. G. MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, Aug. 2016. ISSN 1476-4687. doi: 10.1038/nature19057. URL `https://www.nature.com/articles/nature19057`.

E. B. Lewis. A gene complex controlling segmentation in Drosophila. *Nature*, 276 (5688):565–570, Dec. 1978. ISSN 0028-0836, 1476-4687. doi: 10.1038/276565a0. URL `http://www.nature.com/articles/276565a0`.

C. Leys, M. Delacre, Y. L. Mora, D. Lakens, and C. Ley. How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Psychology*, 32(1):5, Apr. 2019. ISSN 2397-8570. doi: 10.5334/irsp.289. URL http://www.rips-irsp.com/articles/10.5334/irsp.289/.

B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, Aug. 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-323. URL https://doi.org/10.1186/1471-2105-12-323.

X. Li, A. Battle, K. J. Karczewski, Z. Zappala, D. A. Knowles, K. S. Smith, K. R. Kukurba, E. Wu, N. Simon, and S. B. Montgomery. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *American journal of human genetics*, 95(3):245–56, Sept. 2014. ISSN 1537-6605. doi: 10.1016/j.ajhg.2014.08.004. URL http://www.ncbi.nlm.nih.gov/pubmed/25192044.

X. Li, Y. Kim, E. K. Tsang, J. R. Davis, F. N. Damani, C. Chiang, G. T. Hess, Z. Zappala, B. J. Strober, A. J. Scott, A. Li, A. Ganna, M. C. Bassik, J. D. Merker, I. M. Hall, A. Battle, and S. B. Montgomery. The impact of rare variation on gene expression across tissues. *Nature*, 550(7675):239–243, Oct. 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature24267. URL http://www.nature.com/articles/nature24267.

Y. I. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science (New York, N.Y.)*, 352(6285):600–604, Apr. 2016. ISSN 1095-9203. doi: 10.1126/science.aad9417.

Y. I. Li, D. A. Knowles, J. Humphrey, A. N. Barbeira, S. P. Dickinson, H. K. Im, and J. K. Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, 50(1):151–158, Jan. 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-017-0004-9. URL http://www.nature.com/articles/s41588-017-0004-9.

Y. Liao, G. K. Smyth, and W. Shi. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47(8):e47, May 2019. ISSN 1362-4962. doi: 10.1093/nar/gkz114.

R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, Dec. 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://www.nature.com/articles/s41592-018-0229-2.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, Dec. 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

N. López-Bigas, B. Audit, C. Ouzounis, G. Parra, and R. Guigó. Are splicing mutations the most frequent cause of hereditary disease? *FEBS letters*, 579(9):1900–1903, Mar. 2005. ISSN 0014-5793. doi: 10.1016/j.febslet.2005.02.047.

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):469–476, Apr. 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13127. URL `http://www.nature.com/articles/nature13127`.

S. Maddirevula, H. Kuwahara, N. Ewida, H. E. Shamseldin, N. Patel, F. Alzahrani, T. AlSheddi, E. AlObeid, M. Alenazi, H. S. Alsaif, M. Alqahtani, M. AlAli, H. Al Ali, R. Helaby, N. Ibrahim, F. Abdulwahab, M. Hashem, N. Hanna, D. Monies, N. Derar, A. Alsagheir, A. Alhashem, B. Alsaleem, H. Alhebbi, S. Wali, R. Umarov, X. Gao, and F. S. Alkuraya. Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. *Genome Biology*, 21(1):145, June 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02053-9. URL `https://doi.org/10.1186/s13059-020-02053-9`.

P. C. Mahalanobis. On test and measures of group divergence : theoretical formulae. *Journal and Proceedings of Asiatic Society of Bengal*, 26(4):541–588, 1930. URL `http://localhost:8080/xmlui/handle/10263/1639`.

E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470 (7333):198–203, Feb. 2011. ISSN 1476-4687. doi: 10.1038/nature09796. URL `https://www.nature.com/articles/nature09796`.

J. S. Mattick, M. Dinger, N. Schonrock, and M. Cowley. Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing. *Medical Journal of Australia*, 209(5):197–199, Sept. 2018. ISSN 0025-729X, 1326-5377. doi: 10.5694/mja17.01176. URL `https://onlinelibrary.wiley.com/doi/abs/10.5694/mja17.01176`.

J. A. Mayr, T. B. Haack, P. Freisinger, D. Karall, C. Makowski, J. Koch, R. G. Feichtinger, F. A. Zimmermann, B. Rolinski, U. Ahting, T. Meitinger, H. Prokisch, and W. Sperl. Spectrum of combined respiratory chain defects. *Journal of Inherited Metabolic Disease*, 38(4):629–640, July 2015. ISSN 0141-8955, 1573-2665. doi: 10.1007/s10545-015-9831-y. URL `http://doi.wiley.com/10.1007/s10545-015-9831-y`.

V. A. McKusick. Mendelian Inheritance in Man and its online version, OMIM. *American journal of human genetics*, 80(4):588–604, Apr. 2007. ISSN 0002-9297.

doi: 10.1086/514346. URL http://www.sciencedirect.com/science/article/pii/S0002929707611215.

W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17 (1):122, Dec. 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0974-4. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4.

C. Mertes, I. F. Scheller, V. A. Yépez, M. H. Çelik, Y. Liang, L. S. Kremer, M. Gusic, H. Prokisch, and J. Gagneur. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nature Communications*, 12(1):529, Dec. 2021. ISSN 2041-1723. doi: 10.1038/s41467-020-20573-7. URL http://www.nature.com/articles/s41467-020-20573-7.

P. Mohammadi, S. E. Castel, B. B. Cummings, J. Einson, C. Sousa, P. Hoffman, S. Donkervoort, Z. Jiang, P. Mohassel, A. R. Foley, H. E. Wheeler, H. K. Im, C. G. Bonnemann, D. G. MacArthur, and T. Lappalainen. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science*, 366(6463): 351–356, Oct. 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aay0256. URL http://www.sciencemag.org/lookup/doi/10.1126/science.aay0256.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1226. URL https://www.nature.com/articles/nmeth.1226.

D. R. Murdock, H. Dai, L. C. Burrage, J. A. Rosenfeld, S. Ketkar, M. F. Müller, V. A. Yépez, J. Gagneur, P. Liu, S. Chen, M. Jain, G. Zapata, C. A. Bacino, H.-T. Chao, P. Moretti, W. J. Craigen, N. A. Hanchard, and B. Lee. Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. *The Journal of Clinical Investigation*, 131(1), Jan. 2021. ISSN 0021-9738. doi: 10.1172/JCI141500. URL https://www.jci.org/articles/view/141500.

K. Neveling, I. Feenstra, C. Gilissen, L. H. Hoefsloot, E.-J. Kamsteeg, A. R. Mensenkamp, R. J. T. Rodenburg, H. G. Yntema, L. Spruijt, S. Vermeer, T. Rinne, K. L. van Gassen, D. Bodmer, D. Lugtenberg, R. de Reuver, W. Buijsman, R. C. Derks, N. Wieskamp, B. van den Heuvel, M. J. Ligtenberg, H. Kremer, D. A. Koolen, B. P. van de Warrenburg, F. P. Cremers, C. L. Marcelis, J. A. Smeitink, S. B. Wortmann, W. A. van Zelst-Stams, J. A. Veltman, H. G. Brunner, H. Scheffer, and M. R. Nelen. A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases. *Human Mutation*, 34(12):1721–1726, Dec. 2013. ISSN 10597794. doi: 10.1002/humu.22450. URL http://doi.wiley.com/10.1002/humu.22450.

S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad.

Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1):30–35, Jan. 2010. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.499. URL http://www.nature.com/articles/ng.499.

S. Nguengang Wakap, D. M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, and A. Rath. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2):165–173, Feb. 2020. ISSN 1476-5438. doi: 10.1038/s41431-019-0508-0. URL https://www.nature.com/articles/s41431-019-0508-0.

S. Nomura, M. Satoh, T. Fujita, T. Higo, T. Sumida, T. Ko, T. Yamaguchi, T. Tobita, A. T. Naito, M. Ito, K. Fujita, M. Harada, H. Toko, Y. Kobayashi, K. Ito, E. Takimoto, H. Akazawa, H. Morita, H. Aburatani, and I. Komuro. Cardiomyocyte gene programs encoding morphological and functional signatures in cardiac hypertrophy and failure. *Nature Communications*, 9(1):4435, Oct. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06639-7. URL https://www.nature.com/articles/s41467-018-06639-7.

S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, Nov. 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg287. URL https://academic.oup.com/bioinformatics/article/19/16/2088/242445.

M. Pala, Z. Zappala, M. Marongiu, X. Li, J. R. Davis, R. Cusano, F. Crobu, K. R. Kukurba, M. J. Gloudemans, F. Reinier, R. Berutti, M. G. Piras, A. Mulas, M. Zoledziewska, M. Marongiu, E. P. Sorokin, G. T. Hess, K. S. Smith, F. Busonero, A. Maschio, M. Steri, C. Sidore, S. Sanna, E. Fiorillo, M. C. Bassik, S. J. Sawcer, A. Battle, J. Novembre, C. Jones, A. Angius, G. R. Abecasis, D. Schlessinger, F. Cucca, and S. B. Montgomery. Population- and individual-specific regulatory variation in Sardinia. *Nature Genetics*, 49(5):700–707, May 2017. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3840. URL http://www.nature.com/articles/ng.3840.

D. Park, Y. Hoshi, and C. C. Kemp. A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-Based Variational Autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, July 2018a. ISSN 2377-3766. doi: 10.1109/LRA.2018.2801475.

E. Park, Z. Pan, Z. Zhang, L. Lin, and Y. Xing. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *The American Journal of Human Genetics*, 102(1):11–26, Jan. 2018b. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2017.11.002. URL https://www.cell.com/ajhg/abstract/S0002-9297(17)30454-8.

R. Patro, S. M. Mount, and C. Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, May 2014. ISSN 1546-1696. doi: 10.1038/nbt.2862. URL https://www.nature.com/articles/nbt.2862.

R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4): 417–419, Apr. 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4197. URL http://www.nature.com/articles/nmeth.4197.

D. D. Pervouchine, D. G. Knowles, and R. Guigó. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics (Oxford, England)*, 29(2):273–4, Jan. 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts678. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts678.

A. A. Philippakis, D. R. Azzariti, S. Beltran, A. J. Brookes, C. A. Brownstein, M. Brudno, H. G. Brunner, O. J. Buske, K. Carey, C. Doll, S. Dumitriu, S. O. M. Dyke, J. T. d. Dunnen, H. V. Firth, R. A. Gibbs, M. Girdea, M. Gonzalez, M. A. Haendel, A. Hamosh, I. A. Holm, L. Huang, M. E. Hurles, B. Hutton, J. B. Krier, A. Misyura, C. J. Mungall, J. Paschall, B. Paten, P. N. Robinson, F. Schiettecatte, N. L. Sobreira, G. J. Swaminathan, P. E. Taschner, S. F. Terry, N. L. Washington, S. Züchner, K. M. Boycott, and H. L. Rehm. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Human Mutation*, 36(10): 915–921, 2015. ISSN 1098-1004. doi: https://doi.org/10.1002/humu.22858. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.22858.

J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289): 768–772, Apr. 2010. ISSN 0028-0836. doi: 10.1038/nature08872. URL http://www.nature.com/articles/nature08872.

R. D. S. Pitceathly, C. Smith, C. Fratter, C. L. Alston, L. He, K. Craig, E. L. Blakely, J. C. Evans, J. Taylor, Z. Shabbir, M. Deschauer, U. Pohl, M. E. Roberts, M. C. Jackson, C. A. Halfpenny, P. D. Turnpenny, P. W. Lunt, M. G. Hanna, A. M. Schaefer, R. McFarland, R. Horvath, P. F. Chinnery, D. M. Turnbull, J. Poulton, R. W. Taylor, and G. S. Gorman. Adults with RRM2B-related mitochondrial disease have distinct clinical and molecular characteristics. *Brain*, 135(11):3392–3403, Nov. 2012. ISSN 0006-8950. doi: 10.1093/brain/aws231. URL https://doi.org/10.1093/brain/aws231.

F. Piva, M. Giulietti, A. B. Burini, and G. Principato. SpliceAid 2: A database of human splicing factors expression data and RNA target motifs. *Human Mutation*, 33(1):81–85, Jan. 2012. ISSN 10597794. doi: 10.1002/humu.21609. URL http://doi.wiley.com/10.1002/humu.21609.

M. W.-L. Popp and L. E. Maquat. Organizing Principles of Mammalian Nonsense-Mediated mRNA Decay. *Annual Review of Genetics*, 47(1):139–165, Nov. 2013. ISSN 0066-4197, 1545-2948. doi: 10.1146/annurev-genet-111212-133424. URL http://www.annualreviews.org/doi/10.1146/annurev-genet-111212-133424.

G. Poste. Molecular diagnostics: a powerful new component of the healthcare value chain. *Expert Review of Molecular Diagnostics*, 1(1):1–5, Jan. 2014. ISSN 1473-7159. doi: 10.1586/14737159.1.1.1. URL https://doi.org/10.1586/14737159.1.1.1.

A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, and Human Cell Atlas Meeting Participants. The Human Cell Atlas. *eLife*, 6:e27041, Dec. 2017. ISSN 2050-084X. doi: 10.7554/eLife.27041. URL https://doi.org/10.7554/eLife.27041.

S. Rentas, K. S. Rathi, M. Kaur, P. Raman, I. D. Krantz, M. Sarmady, and A. A. Tayoun. Diagnosing Cornelia de Lange syndrome and related neurodevelopmental disorders using RNA sequencing. *Genetics in Medicine*, 22(5):927–936, May 2020. ISSN 1530-0366. doi: 10.1038/s41436-019-0741-5. URL https://www.nature.com/articles/s41436-019-0741-5.

K. Retterer, J. Juusola, M. T. Cho, P. Vitazka, F. Millan, F. Gibellini, A. Vertino-Bell, N. Smaoui, J. Neidich, K. G. Monaghan, D. McKnight, R. Bai, S. Suchy, B. Friedman, J. Tahiliani, D. Pineda-Alvarez, G. Richard, T. Brandt, E. Haverfield, W. K. Chung, and S. Bale. Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine*, 18(7):696–704, July 2016. ISSN 1098-3600, 1530-0366. doi: 10.1038/gim.2015.148. URL http://www.nature.com/articles/gim2015148.

J. L. Reyes, P. Kois, B. B. Konforti, and M. M. Konarska. The canonical GU dinucleotide at the 5' splice site is recognized by p220 of the U5 snRNP within the spliceosome. *RNA*, 2(3):213–225, Mar. 1996. ISSN 1355-8382. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1369364/.

S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. 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, 1530-0366. doi: 10.1038/gim.2015.30. URL http://www.nature.com/articles/gim201530.

T. Richter, S. Nestler-Parr, R. Babela, Z. M. Khan, T. Tesoro, E. Molsen, and D. A. Hughes. Rare Disease Terminology and Definitions—A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value in Health*, 18 (6):906–914, Sept. 2015. ISSN 1098-3015, 1524-4733. doi: 10.1016/j.jval.2015.

05.008. URL https://www.valueinhealthjournal.com/article/S1098-3015(15)01979-8/abstract.

D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*, 12(1):480, 2011. ISSN 14712105. doi: 10.1186/1471-2105-12-480. URL http://www.biomedcentral.com/1471-2105/12/480.

A. Roberts and L. Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, Jan. 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2251. URL https://www.nature.com/articles/nmeth.2251.

J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):3, 2011.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–40, Jan. 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp616. URL https://doi.org/10.1093/bioinformatics/btp616.

J. Rode. Rare Diseases: Understanding this Public Health Priority. *Rare Diseases*, page 14, 2005. URL https://www.eurordis.org.

M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, 242 (1):84–89, Nov. 1996. ISSN 0003-2697. doi: 10.1006/abio.1996.0432. URL https://www.sciencedirect.com/science/article/pii/S0003269796904327.

A. B. Rosenberg, R. P. Patwardhan, J. Shendure, and G. Seelig. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, 163(3):698–711, Oct. 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.09.054.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, Dec. 1977. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.74.12.5463. URL http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463.

Z. E. Sauna and C. Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10):683–691, Oct. 2011. ISSN 1471-0064. doi: 10.1038/nrg3051. URL https://www.nature.com/articles/nrg3051.

M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270:467–470, Oct. 1995.

M. M. Scotti and M. S. Swanson. RNA mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19–32, Nov. 2015. ISSN 1471-0056. doi: 10.1038/nrg.2015.3. URL http://www.nature.com/doifinder/10.1038/nrg.2015.3.

A. L. Shaffer, W. Wojnar, and W. Nelson. Amplification, detection, and automated sequencing of gibbon interleukin-2 mRNA by Thermus aquaticus DNA polymerase reverse transcription and polymerase chain reaction. *Analytical Biochemistry*, 190(2): 292–296, Nov. 1990. ISSN 0003-2697. doi: 10.1016/0003-2697(90)90196-G. URL https://www.sciencedirect.com/science/article/pii/000326979090196G.

H. E. Shamseldin, S. Maddirevula, E. Faqeih, N. Ibrahim, M. Hashem, R. Shaheen, and F. S. Alkuraya. Increasing the sensitivity of clinical exome sequencing through improved filtration strategy. *Genetics in Medicine*, 19(5):593–598, May 2017. ISSN 1098-3600, 1530-0366. doi: 10.1038/gim.2016.155. URL http://www.nature.com/articles/gim2016155.

S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601, Dec. 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1419161111. URL https://www.pnas.org/content/111/51/E5593.

C. R. Sibley, L. Blazquez, and J. Ule. Lessons from non-canonical splicing. *Nature Publishing Group*, 17, 2016. doi: 10.1038/nrg.2016.46.

B. Sikkema-Raddatz, L. F. Johansson, E. N. de Boer, R. Almomani, L. G. Boven, M. P. van den Berg, K. Y. van Spaendonck-Zwarts, J. P. van Tintelen, R. H. Sijmons, J. D. H. Jongbloed, and R. J. Sinke. Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics. *Human Mutation*, 34(7):1035–1042, July 2013. ISSN 10597794. doi: 10.1002/humu.22332. URL http://doi.wiley.com/10.1002/humu.22332.

W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9): 1164–1167, May 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm069. URL https://academic.oup.com/bioinformatics/article/23/9/1164/272597.

O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, Mar. 2012. ISSN 1750-2799. doi: 10.1038/nprot.2011.457. URL https://www.nature.com/articles/nprot.2011.457.

M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891):956–960, Aug. 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1160342. URL https://science.sciencemag.org/content/321/5891/956.

M. Sun. Mucolipidosis type IV is caused by mutations in a gene encoding a novel transient receptor potential channel. *Human Molecular Genetics*, 9(17):2471–2478,

Oct. 2000. ISSN 14602083. doi: 10.1093/hmg/9.17.2471. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/9.17.2471.

G. J. Swaminathan, E. Bragin, E. A. Chatzimichali, M. Corpas, A. P. Bevan, C. F. Wright, N. P. Carter, M. E. Hurles, and H. V. Firth. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Human Molecular Genetics*, 21(R1):R37–R44, Oct. 2012. ISSN 0964-6906. doi: 10.1093/hmg/dds362. URL https://doi.org/10.1093/hmg/dds362.

K. Szczepanowska, P. Maiti, A. Kukat, E. Hofsetz, H. Nolte, K. Senft, C. Becker, B. Ruzzenente, H. Hornig-Do, R. Wibom, R. J. Wiesner, M. Krüger, and A. Trifunovic. CLPP coordinates mitoribosomal assembly through the regulation of ERAL1 levels. *The EMBO Journal*, 35(23):2566–2583, Dec. 2016. ISSN 0261-4189, 1460-2075. doi: 10.15252/embj.201694253. URL https://onlinelibrary.wiley.com/doi/10.15252/embj.201694253.

J. C. Taylor, H. C. Martin, S. Lise, J. Broxholme, J.-B. Cazier, A. Rimmer, A. Kanapin, G. Lunter, S. Fiddy, C. Allan, A. R. Aricescu, M. Attar, C. Babbs, J. Becq, D. Beeson, C. Bento, P. Bignell, E. Blair, V. J. Buckle, K. Bull, O. Cais, H. Cario, H. Chapel, R. R. Copley, R. Cornall, J. Craft, K. Dahan, E. E. Davenport, C. Dendrou, O. Devuyst, A. L. Fenwick, J. Flint, L. Fugger, R. D. Gilbert, A. Goriely, A. Green, I. H. Greger, R. Grocock, A. V. Gruszczyk, R. Hastings, E. Hatton, D. Higgs, A. Hill, C. Holmes, M. Howard, L. Hughes, P. Humburg, D. Johnson, F. Karpe, Z. Kingsbury, U. Kini, J. C. Knight, J. Krohn, S. Lamble, C. Langman, L. Lonie, J. Luck, D. McCarthy, S. J. McGowan, M. F. McMullin, K. A. Miller, L. Murray, A. H. Németh, M. A. Nesbit, D. Nutt, E. Ormondroyd, A. B. Oturai, A. Pagnamenta, S. Y. Patel, M. Percy, N. Petousi, P. Piazza, S. E. Piret, G. Polanco-Echeverry, N. Popitsch, F. Powrie, C. Pugh, L. Quek, P. A. Robbins, K. Robson, A. Russo, N. Sahgal, P. A. van Schouwenburg, A. Schuh, E. Silverman, A. Simmons, P. S. Sørensen, E. Sweeney, J. Taylor, R. V. Thakker, I. Tomlinson, A. Trebes, S. R. F. Twigg, H. H. Uhlig, P. Vyas, T. Vyse, S. A. Wall, H. Watkins, M. P. Whyte, L. Witty, B. Wright, C. Yau, D. Buck, S. Humphray, P. J. Ratcliffe, J. I. Bell, A. O. M. Wilkie, D. Bentley, P. Donnelly, and G. McVean. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, 47(7):717–726, July 2015. ISSN 1546-1718. doi: 10.1038/ng.3304. URL https://www.nature.com/articles/ng.3304.

K. Taylor and K. Sobczak. Intrinsic Regulatory Role of RNA Structural Arrangement in Alternative Splicing Control. *International Journal of Molecular Sciences*, 21(14):5161, July 2020. ISSN 1422-0067. doi: 10.3390/ijms21145161. URL https://www.mdpi.com/1422-0067/21/14/5161.

R. C. Team. R: A Language and Environment for Statistical Computing, 2021. URL https://www.R-project.org/.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct. 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature15393. URL `http://www.nature.com/articles/nature15393`.

The GTEx Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235): 648–660, May 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1262110. URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1262110`.

R. Thompson, L. Johnston, D. Taruscio, L. Monaco, C. Béroud, I. G. Gut, M. G. Hansson, P.-B. A. 't Hoen, G. P. Patrinos, H. Dawkins, M. Ensini, K. Zatloukal, D. Koubi, E. Heslop, J. E. Paschall, M. Posada, P. N. Robinson, K. Bushby, and H. Lochmüller. RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research. *Journal of General Internal Medicine*, 29(3):780–787, Aug. 2014. ISSN 1525-1497. doi: 10.1007/s11606-014-2908-8. URL `https://doi.org/10.1007/s11606-014-2908-8`.

H. Tilgner, D. G. Knowles, R. Johnson, C. A. Davis, S. Chakrabortty, S. Djebali, J. Curado, M. Snyder, T. R. Gingeras, and R. Guigó. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625, Sept. 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.134445.111. URL `http://genome.cshlp.org/content/22/9/1616`.

M. J. T. N. Timmermans, S. Dodsworth, C. L. Culverwell, L. Bocak, D. Ahrens, D. T. J. Littlewood, J. Pons, and A. P. Vogler. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*, 38(21):e197–e197, Nov. 2010. ISSN 1362-4962, 0305-1048. doi: 10.1093/nar/gkq807. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq807.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010. ISSN 1546-1696. doi: 10.1038/nbt.1621. URL https://www.nature.com/articles/nbt.1621.

J. L. Trincado, J. C. Entizne, G. Hysenaj, B. Singh, M. Skalic, D. J. Elliott, and E. Eyras. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19(1):40, Mar. 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1417-1. URL https://doi.org/10.1186/s13059-018-1417-1.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, June 2001. ISSN 1367-4803. doi: 10.1093/bioinformatics/17.6.520. URL https://academic.oup.com/bioinformatics/article/17/6/520/272365.

S. B. Vafai and V. K. Mootha. Mitochondrial disorders as windows into an ancient organelle. *Nature*, 491(7424):374–383, Nov. 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature11707. URL http://www.nature.com/articles/nature11707.

B. van de Geijn, G. McVicker, Y. Gilad, and J. K. Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, 12 (11):1061–1063, Nov. 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3582. URL https://www.nature.com/articles/nmeth.3582.

K. Van Keuren-Jensen, J. J. Keats, and D. W. Craig. Bringing RNA-seq closer to the clinic. *Nature Biotechnology*, 32(9):884–885, Sept. 2014. ISSN 1546-1696. doi: 10.1038/nbt.3017. URL https://www.nature.com/articles/nbt.3017.

J. Vaquero-Garcia, A. Barrera, M. R. Gazzara, J. González-Vallinas, N. F. Lahens, J. B. Hogenesch, K. W. Lynch, and Y. Barash. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, 5:e11752, Feb. 2016. ISSN 2050-084X. doi: 10.7554/eLife.11752. URL https://doi.org/10.7554/eLife.11752.

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1096–1103, Helsinki, Finland,

2008. ACM Press. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294. URL http://portal.acm.org/citation.cfm?doid=1390156.1390294.

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *The Journal of Machine Learning Research*, 11:3371–3408, Dec. 2010. ISSN 1532-4435.

G.-S. Wang and T. A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews. Genetics*, 8(10):749–761, Oct. 2007. ISSN 1471-0064. doi: 10.1038/nrg2164.

K. Wang, C. Kim, J. Bradfield, Y. Guo, E. Toskala, F. G. Otieno, C. Hou, K. Thomas, C. Cardinale, G. J. Lyon, R. Golhar, and H. Hakonarson. Whole-genome DNA/RNA sequencing identifies truncating mutations in RBCK1 in a novel Mendelian disease with neuromuscular and cardiac involvement. *Genome Medicine*, 5(7):67, 2013. ISSN 1756-994X. doi: 10.1186/gm471. URL http://genomemedicine.biomedcentral.com/articles/10.1186/gm471.

Z. Wang and C. B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, May 2008. ISSN 1469-9001. doi: 10.1261/rna.876308.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, Jan. 2009. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2484. URL http://www.nature.com/articles/nrg2484.

J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, Apr. 1953. ISSN 1476-4687. doi: 10.1038/171737a0. URL https://www.nature.com/articles/171737a0.

G. P. Way and C. S. Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 23:80–91, 2018. ISSN 2335-6936.

B.-J. M. Webb-Robertson, K. G. Stratton, J. E. Kyle, Y.-M. Kim, L. M. Bramer, K. M. Waters, D. M. Koeller, and T. O. Metz. Statistically Driven Metabolite and Lipid Profiling of Patients from the Undiagnosed Diseases Network. *Analytical Chemistry*, 92(2):1796–1803, Jan. 2020. ISSN 0003-2700. doi: 10.1021/acs.analchem.9b03522. URL https://doi.org/10.1021/acs.analchem.9b03522.

A. P. Weber. Discovering New Biology through Sequencing of RNA. *Plant Physiology*, 169(3):1524–1531, Nov. 2015. ISSN 0032-0889. doi: 10.1104/pp.15.01081. URL https://doi.org/10.1104/pp.15.01081.

A. P. M. Weber, K. L. Weber, K. Carr, C. Wilkerson, and J. B. Ohlrogge. Sampling the Arabidopsis Transcriptome with Massively Parallel Pyrosequencing. *Plant Physiology*,

144(1):32–42, May 2007. ISSN 0032-0889, 1532-2548. doi: 10.1104/pp.107.096677. URL http://www.plantphysiol.org/content/144/1/32.

L. Whitaker. On the Poisson Law of Small Numbers. *Biometrika*, 10(1):36–71, 1914. ISSN 0006-3444. doi: 10.2307/2331739. URL https://www.jstor.org/stable/2331739.

M. Witzel, D. Petersheim, Y. Fan, E. Bahrami, T. Racek, M. Rohlfs, J. Puchałka, C. Mertes, J. Gagneur, C. Ziegenhain, W. Enard, A. Stray-Pedersen, P. Arkwright, M. Abboud, V. Pazhakh, G. Lieschke, P. Krawitz, M. Dahlhoff, M. Schneider, E. Wolf, H.-P. Horny, H. Schmidt, A. Schäffer, and C. Klein. Chromatin-remodeling factor SMARCD2 regulates transcriptional networks controlling differentiation of neutrophil granulocytes. *Nature Genetics*, 49(5), 2017. ISSN 15461718. doi: 10.1038/ng.3833.

H. Wold. *Estimation of principal components and related models by iterative least squares*. Academic Press, NewYork, 1966.

S. Wortmann, J. Mayr, J. Nuoffer, H. Prokisch, and W. Sperl. A Guideline for the Diagnosis of Pediatric Mitochondrial Disease: The Value of Muscle and Skin Biopsies in the Genetics Era. *Neuropediatrics*, 48(04):309–314, Aug. 2017. ISSN 0174-304X, 1439-1899. doi: 10.1055/s-0037-1603776. URL http://www.thieme-connect.de/DOI/DOI?10.1055/s-0037-1603776.

S. B. Wortmann, D. A. Koolen, J. A. Smeitink, L. van den Heuvel, and R. J. Rodenburg. Whole exome sequencing of suspected mitochondrial patients in clinical practice. *Journal of Inherited Metabolic Disease*, 38(3):437–443, May 2015. ISSN 0141-8955, 1573-2665. doi: 10.1007/s10545-015-9823-y. URL http://doi.wiley.com/10.1007/s10545-015-9823-y.

C. F. Wright, D. R. FitzPatrick, and H. V. Firth. Paediatric genomics: diagnosing rare disease in children. *Nature Reviews Genetics*, 19(5):253–268, May 2018a. ISSN 1471-0064. doi: 10.1038/nrg.2017.116. URL https://www.nature.com/articles/nrg.2017.116.

C. F. Wright, J. F. McRae, S. Clayton, G. Gallone, S. Aitken, T. W. FitzGerald, P. Jones, E. Prigmore, D. Rajan, J. Lord, A. Sifrim, R. Kelsell, M. J. Parker, J. C. Barrett, M. E. Hurles, D. R. FitzPatrick, H. V. Firth, and DDD Study. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genetics in Medicine*, 20(10):1216–1223, Oct. 2018b. ISSN 1098-3600, 1530-0366. doi: 10.1038/gim.2017.246. URL http://www.nature.com/articles/gim2017246.

H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey. The human splicing code reveals new insights into the genetic determinants of disease. *Science (New York,*

N.Y.), 347(6218):1254806, Jan. 2015. ISSN 0036-8075. doi: 10.1126/science.1254806. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4362528/.

T. W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer Series in Statistics. Springer New York, New York, NY, 2015. ISBN 978-1-4939-2818-7. doi: 10.1007/978-1-4939-2818-7_1. URL https://doi.org/10.1007/978-1-4939-2818-7_1.

G. Yeo, D. Holste, G. Kreiman, and C. B. Burge. Variation in alternative splicing across human tissues. *Genome Biology*, 5(10):R74, Sept. 2004a. ISSN 1474-760X. doi: 10.1186/gb-2004-5-10-r74. URL https://doi.org/10.1186/gb-2004-5-10-r74.

G. Yeo, S. Hoon, B. Venkatesh, and C. B. Burge. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proceedings of the National Academy of Sciences*, 101(44):15700–15705, Nov. 2004b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0404901101. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0404901101.

J. Yu, C. Hong, Y. Rui, and D. Tao. Multitask Autoencoder Model for Recovering Human Poses. *IEEE Transactions on Industrial Electronics*, 65(6):5060–5068, June 2018. ISSN 1557-9948. doi: 10.1109/TIE.2017.2739691.

V. A. Yépez, M. Gusic, R. Kopajtich, C. Mertes, N. H. Smith, C. L. Alston, R. Berutti, H. Blessing, E. Ciara, F. Fang, P. Freisinger, D. Ghezzi, S. J. Hayflick, Y. Kishita, T. Klopstock, C. Lamperti, D. Lenz, C. C. Makowski, J. A. Mayr, S. Mosegaard, M. F. Müller, G. Muñoz-Pujol, K. Murayama, A. Nadel, A. Ohtake, Y. Okazaki, D. Piekutowska-Abramczuk, E. Procopio, A. Ribes, A. Rötig, T. Schwarzmayr, C. Staufner, S. L. Stenton, T. M. Strom, R. W. Taylor, C. Terrile, F. Tort, R. V. Coster, M. Wagner, S. B. Wortmann, M. Xu, T. Meitinger, J. Gagneur, and H. Prokisch. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *medRxiv*, page 2021.04.01.21254633, Apr. 2021a. doi: 10.1101/2021.04.01.21254633. URL https://www.medrxiv.org/content/10.1101/2021.04.01.21254633v1.

V. A. Yépez, C. Mertes, M. F. Müller, D. Klaproth-Andrade, L. Wachutka, L. Frésard, M. Gusic, I. F. Scheller, P. F. Goldberg, H. Prokisch, and J. Gagneur. Detection of aberrant gene expression events in RNA sequencing data. *Nature Protocols*, 16(2):1276–1296, Feb. 2021b. ISSN 1750-2799. doi: 10.1038/s41596-020-00462-5. URL https://www.nature.com/articles/s41596-020-00462-5.

D. T. Zallen. Despite Franklin's work, Wilkins earned his Nobel. *Nature*, 425(6953):15–15, Sept. 2003. ISSN 1476-4687. doi: 10.1038/425015b. URL https://www.nature.com/articles/425015b.

Y. Zeng, G. Wang, E. Yang, G. Ji, C. L. Brinkmeyer-Langford, and J. J. Cai. Aberrant Gene Expression in Humans. *PLOS Genetics*, 11(1):e1004942, Jan. 2015. ISSN 1553-

7404. doi: 10.1371/journal.pgen.1004942. URL https://dx.plos.org/10.1371/journal.pgen.1004942.

Z. Zeng and Y. Bromberg. Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00914. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00914/full.

X. Zhou, H. Lindsay, and M. D. Robinson. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research*, 42(11): e91–e91, June 2014. ISSN 0305-1048. doi: 10.1093/nar/gku310. URL https://doi.org/10.1093/nar/gku310.

A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *WIREs Data Mining and Knowledge Discovery*, 8(6):e1280, 2018. ISSN 1942-4795. doi: https://doi.org/10.1002/widm.1280. URL https://www.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1280.