TUM School of Computation, Information and Technology
Technische Universität München

Aberrant gene expression prediction from sequence and implications in health and disease

Florian Rupert Hölzlwimmer

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Nassir Navab

Prüfende der Dissertation:

1. Prof. Dr. Julien Gagneur
2. Prof. Dr. Bertram Müller-Myhsok

Die Dissertation wurde am 22.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 26.09.2024 angenommen.

Acknowledgements

First and foremost, I am deeply grateful to Prof. Julien Gagneur for his unwavering support, patience, and invaluable guidance. You always had an open door and advised me, even during your tightest schedule. Your guidance helped me not only with my own research work but also taught me to be a better supervisor to others. You created a lab with an awesome atmosphere and a collaborative environment in which I very much enjoyed to work. Next, I want to thank my mentor, Dr. Vicente Yépez, not only for providing me with advice and support, but also for being a great friend during the highs and lows of the last five years. I would like to express my gratitude to Felix Brechtmann and Alexander Karollus for the long and fruitful discussions from which so many great ideas have emerged. A big thank you goes to Nils Wagner and Muhammed Hasan Çelik for their dedicated work and valuable feedback on aberrant splicing. I would also like to thank Jonas Lindner for his great work and contribution on the UK Biobank study as well as Eva Holtkamp for her feedback. I am also very grateful for the constructive feedback provided by my thesis advisory committee, Prof. Annalisa Marsico and Dr. Francesco Paolo Casale. Especially Paolos continuous advice over the last two years was highly appreciated.

Next, I want to thank everybody in the IT administration team, Vanda, Johannes, Thomas, and Chris, for doing an outstanding job and allowing me to focus on my research. I especially want to thank Vanda, without her constant help over the last years I would not have finished this dissertation today. Also, I would like to express my appreciation to Andrea Wolf and Inga Weise for their consistent assistance with administrative matters.

I want to thank all my former and current colleagues from the Gagneurlab, where I did not only find colleagues but friends. Thank you for all the parties, movie nights, and journeys together. Especially I want to thank Daniela, for your high spirits and an unforgettable trip to Ecuador. Pedro, for all the cake and your guitar play. Ines and Mada for all the great years of studying together and introducig me to the lab. Vangelis, for all the fun facts and jokes. Xueqi, for being a wonderful and attentive office mate. Leo, for your awesome Feuerzangenbowle and Kaiserschmarrn. Chris, for your support and encouragement. Additionally, I want to thank all the students I supervised, including Paul, Timon, Jonas, and Julian.

Ein besonderer Dank gilt meinen Eltern Rupert und Irene für die jahrelange Unterstützung, ebenso wie meinen Geschwistern Adrian und Stefanie sowie meiner Großmutter Irene. Ohne euch würde ich heute nicht diese Zeilen schreiben. Außerdem möchte ich Stephan für die vielen Jahre der Freundschaft und das Korrekturlesen meiner Arbeit danken. Last but not least, my deepest appreciation goes to my partner Vanda, who is always on my side, encouraging me, and keeping my life balanced.

Abstract

Identifying genetic variants with high functional impact is essential for deciphering the genetic underpinning of diseases. Despite the important role of aberrant gene expression in diseases, the impact of genetic variants on gene expression in different tissues often remains unknown. Detecting aberrantly expressed genes in RNA sequencing data of affected tissues can aid in identifying disruptive large-effect variants. However, this approach is limited to clinically accessible tissues such as skin or body fluids and does not generalize to unseen variants.

Here I set out to predict rare variants associated with aberrant underexpression across 48 human tissues. To achieve this, I established the first systematic benchmark for expression outlier prediction by applying OUTRIDER, an aberrant expression caller using RNA-seq count data, to 11,096 GTEx samples. I assessed and developed predictors that use DNA sequence and optionally RNA sequencing data from clinically accessible tissues. Although not developed for this task, existing methods such as LOFTEE, CADD, and AbSplice-DNA exhibited mild predictive performance in predicting aberrantly underexpressed genes (0.5-1.5% average precision in median across tissue types).

Building on these results, I aimed to improve the prediction of underexpression outliers. Therefore, I developed AbExp, a specialized tool for aberrant underexpression across human tissues that takes DNA sequence as input and predicts a continuous, tissue-specific z-score of gene expression. By integrating various variant effect annotations with the proportion of affected isoforms per tissue as well as considering tissue-specific gene expression variability, AbExp reaches an average precision of 9.1% in median across tissue types, outperforming existing tools between 6-fold and 18-fold. Testing AbExp on independent datasets confirmed the consistency of the performance improvements and permitted the differentiation between pathogenic and benign variants with high precision. Integrating AbExp predictions with gene expression measurements from clinically accessible tissues yielded another two-fold enhancement in predicting tissue-specific aberrant expression in non-accessible tissues.

Finally, I demonstrated how AbExp can improve rare variant gene association testing as well as phenotype prediction on 40 blood traits from 200,000 individuals of the UK Biobank. An AbExp-based association test identified 30% more trait-associated genes compared to a LOFTEE-based burden test. In addition, AbExp scores significantly improved phenotype prediction over LOFTEE in 50% of the traits, while never exhibiting inferior performance.

In summary, the development of a DNA-sequence-based method for predicting aberrant gene expression in multiple tissues, which can also generalize to unseen variants, represents a significant advancement in the ability to identify and understand the genetic underpinnings of human traits and diseases.

Publications

Aberrant Expression Prediction across Human Tissues

Florian R. Hölzlwimmer, Jonas Lindner, Nils Wagner, Francesco Paolo Casale, Vicente A. Yépez, Julien Gagneur

(2023) **bioRxiv**, DOI: 10.1101/2023.12.04.569414, Ref: [61]

Author contribution:
J.G. conceptualized the project. F.R.H., F.P.C. and J.G. designed the methodology. F.R.H., N.W. and V.A.Y. curated the data. F.R.H. and J.L. performed investigation and provided the software. F.R.H., J.L., and V.A.Y.. performed visualizations. F.R.H., V.A.Y. and J.G. wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. J.G. supervised the project with the help of F.P.C. and V.A.Y.

Additionally, I contributed to the following papers:

Aberrant Splicing Prediction across Human Tissues

Nils Wagner*, Muhammed H. Çelik*, **Florian R. Hölzlwimmer**, Christian Mertes, Holger Prokisch, Vicente A. Yépez, Julien Gagneur

* Equal contribution

(2023) **Nature Genetics**, DOI: 10.1038/s41588-023-01373-3, Ref: [144]

Author contribution:
J.G. conceptualized the project. N.W., M.H.C. and J.G. designed the methodology. N.W. and M.H.C. provided the software. N.W., M.H.C., F.R.H., H.P. and V.A.Y. performed validations. N.W., M.H.C., F.R.H., V.A.Y. and C.M. performed the formal analysis. N.W., M.H.C., F.R.H. and V.A.Y. curated the data. N.W., M.H.C., V.A.Y. and J.G. wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. N.W., M.H.C., F.R.H., V.A.Y. and J.G. performed visualizations. J.G. supervised the project.

Integration of Variant Annotations Using Deep Set Networks Boosts Rare Variant Association Genetics

Brian Clarke*, Eva Holtkamp*, Hakime Öztürk, Marcel Mück, Magnus Wahlberg, Kayla Meyer, Felix Munzlinger, Felix Brechtmann, **Florian R. Hölzlwimmer**, Julien Gagneur, Oliver Stegle

* Equal contribution

(2023) **bioRxiv** , DOI: 10.1101/2023.07.12.548506, Ref: [22]

Author contribution:
B.C., E.H., J.G., & O.S. conceived the method. F.B. made additional conceptual contributions to the method. B.C., E.H., H.O., & F.H. prepared the data. B.C., E.H., H.O., K.M, M.M., F.M., & M.W. implemented the methods and analyzed the data. B.C., E.H., J.G., & O.S. interpreted the results and wrote the paper.

Contents

Acknowledgements **iii**

Abstract **v**

Publications **vii**

Contents **ix**

1 Introduction **1**
1.1 Understanding variant impact is key to studying human disease 1
1.2 Most high-impact variants are rare 4
1.3 Biology of high-impact variants 8
1.3.1 Coding variants 8
1.3.2 Regulatory variants 8
1.3.3 Integrating genomics with transcriptomics to identify high-impact variants 11
1.4 Aims and scope of this thesis 13
1.4.1 Benchmarking aberrant gene expression prediction in human tissues 13
1.4.2 AbExp: Predicting aberrant gene underexpression across human tissues 13
1.4.3 Improving rare variant association testing and phenotype prediction with AbExp 13

2 Background **15**
2.1 Genetic variants 15
2.1.1 Types of genetic variants 15
2.1.2 Zygosity of variants 15
2.2 Central dogma of biology 15
2.3 Nonsense-mediated decay of mRNA 17
2.4 Next-Generation Sequencing 20
2.4.1 DNA-seq 21
2.4.2 RNA-seq 22
2.5 OUTRIDER: Aberrant gene expression calling 23
2.6 Precision-Recall Curve and Average Precision 25

2.7 Supervised learning . . . 25
2.7.1 Linear regression . . . 26
2.7.2 Logistic regression . . . 28
2.7.3 Elastic Net Regularization . . . 28
2.7.4 LightGBM: Gradient boosting decision trees . . . 30
2.8 Likelihood Ratio Test . . . 32
2.9 Ensembl VEP . . . 35
2.10 LOFTEE . . . 35
2.11 Combined Annotation Dependent Depletion (CADD) . . . 36
2.12 Datasets . . . 36
2.12.1 GTEx . . . 36
2.12.2 Answer ALS . . . 38
2.12.3 Mitochondrial disease dataset . . . 38
2.12.4 UK Biobank . . . 39
2.12.5 ClinVar . . . 39
2.12.6 GnomAD . . . 39

3 Benchmarking aberrant gene expression prediction in human tissues 41
3.1 Motivation . . . 41
3.2 A benchmark for tissue-specific aberrant expression prediction . . . 41
3.2.1 Expression outlier calling . . . 43
3.2.2 Filtering of expression outliers . . . 43
3.2.3 Rare variant filtering . . . 45
3.2.4 Benchmark task and evaluation metric . . . 51
3.3 Performance of variant annotation tools in aberrant expression prediction 51
3.3.1 Enrichment of variant consequences . . . 51
3.3.2 LOFTEE . . . 53
3.3.3 AbSplice . . . 53
3.3.4 CADD . . . 56
3.3.5 Performance comparison of LOFTEE, AbSplice, and CADD on aberrant underexpression prediction . . . 56
3.4 Summary . . . 56

4 AbExp: Predicting aberrant gene underexpression across human tissues 59
4.1 Motivation . . . 59
4.2 Training and evaluation procedure . . . 59
4.3 Integrating rare variant annotations to predict underexpression outliers across tissues . . . 59
4.3.1 Calculation of gene-level features . . . 61
4.3.2 Quantitative prediction of outlier state . . . 61

4.4 Accounting for tissue-specific isoform expression . . . 61
4.4.1 Calculation of transcript isoform proportions in each tissue . . . 65
4.4.2 Calculation of gene-level features . . . 65
4.5 Incorporating the tissue-specific gene expression variability . . . 68
4.6 Contribution of aberrant splicing and transcript ablations . . . 70
4.7 AbExp performance replicates on independent datasets . . . 75
4.7.1 Outlier calling and rare variant filtering . . . 75
4.7.2 Performance evaluation . . . 75
4.8 Analysis of AbExp scores . . . 79
4.8.1 AbExp predicts on average 1.2 high-confidence and 5.7 low-confidence underexpressed genes per individual . . . 79
4.8.2 AbExp predictions are tissue-specific . . . 79
4.8.3 25-45% of AbExp high-confidence predictions can not be explained with LOFTEE . . . 79
4.8.4 AbExp correlation with measured expression varies . . . 82
4.8.5 AbExp predicts pathogenic variants with high precision . . . 84
4.9 The AbExp variant effect prediction pipeline . . . 84
4.10 Integrating AbExp with gene expression measurements from clinically accessible tissues . . . 86
4.11 Summary . . . 88

5 Improving rare variant association testing and phenotype prediction with AbExp 91
5.1 Motivation . . . 91
5.2 Rare variant association testing . . . 91
5.2.1 UK Biobank genome and phenotype data . . . 93
5.2.2 Identification of lead trait-associated common variants . . . 96
5.2.3 Application of polygenic risk scores . . . 96
5.2.4 Variant filtering and annotation . . . 96
5.2.5 P-value calculation and calibration . . . 96
5.3 Phenotype prediction . . . 99
5.3.1 Model training and evaluation . . . 99
5.3.2 AbExp affects the prediction of extreme phenotypes . . . 99
5.4 Summary . . . 102

6 Discussion 103
6.1 Benchmarking aberrant gene expression prediction in human tissues . . . 103
6.2 AbExp: Predicting aberrant gene underexpression across human tissues . 104
6.2.1 AbExp assumes that outliers are caused by rare variants within gene regions . . . 104
6.2.2 Additional annotations could improve AbExp . . . 105
6.2.3 Aberrant expression prediction in more settings . . . 109

6.3 The UK Biobank rare variant association testing and phenotype prediction study can be improved 110
6.4 Conclusion and outlook 111

A Data and code availability **113**

B Supplementary figures **115**

List of Figures **119**

List of Tables **123**

Acronyms **125**

References **127**

1 Introduction

The human genome is the blueprint for our lives. It contains all the necessary instructions on how we develop from a single fertilized egg cell into a complex organism consisting of trillions of specialized cells. Our genome consists of 23 pairs of homologous nuclear chromosomes and one set of mitochondrial DNA. Among nuclear chromosomes, one pair consists of sex chromosomes, while the remaining 22 pairs are comprised of sex-unspecific chromosomes, also called autosomes. The mitochondrial DNA and one copy of each chromosome, including 22 autosomes and an X chromosome, are inherited from our mother. The other half of chromosomes, including 22 autosomes and either an X or a Y chromosome which determines sex, is inherited from our father. Each chromosome is a packed, double-stranded molecule of deoxyribonucleic acid (DNA) representing a sequence of cytosine (C), guanine (G), adenine (A), and thymine (T) nucleobases. Encoded in a total sequence length of about 6 billion nucleobases, the DNA contains a set of genes as well as regulatory elements that control when and how much each gene is being expressed[3, 112]. While 99% of the DNA sequence is shared across humans, every one of us has about 4-5 million variations in this sequence that distinguish us from the average population[5] (see section 2.1). Genetic variants determine our individual traits such as the color of our hair and eyes, whether we can digest dairy products as adults, and even whether we enjoy drinking coffee[64, 103, 137].

1.1 Understanding variant impact is key to studying human disease

Embedded within the DNA are also variations that affect or cause our susceptibility to many diseases. Pinpointing these variants and understanding their biological impact is an ongoing effort in genomic medicine for several reasons. First of all, understanding one's genetic predisposition allows for preventive measures and more tailored treatments[50, 142]. By identifying who is at higher risk for certain diseases, healthcare providers can recommend proactive lifestyle changes, such as dietary adjustments or physical exercise, and more frequent screening for early detection. For example, testing for high-impact variants in cancer driver genes, such as *BRCA1* and *BRCA2*, can identify individuals at high risk for certain types of cancer [42, 99]. For such persons, regular cancer screening can significantly enhance treatment efficacy and long-term prospects of survival.

Genetic testing is also useful in the context of in-vitro fertilization. During the preim-

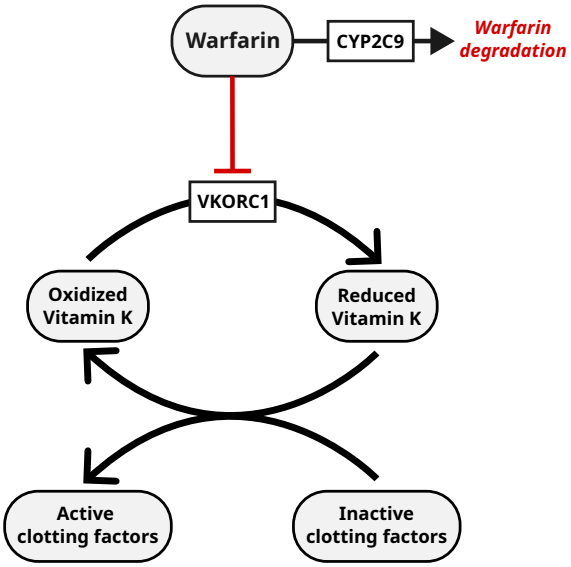

Figure 1.1: Warfarin mechanism of action. Warfarin works by inhibiting the activity of a vitamin K-producing enzyme complex encoded by the *VKORC1* gene. This inhibition in turn reduces the activity of several blood clotting factors that depend on vitamin K for their activation[113]. Additionally, Warfarin is metabolized in the liver by enzymes, one of which is encoded by the *CYP2C9* gene. Therefore, variants in the *VKORC1* and *CYP2C9* genes can affect the concentration and efficacy of Warfarin[87].

plantation stage, embryos are tested for pathogenic (i.e. disease-causing) variations. This testing is crucial for selecting embryos that are most viable and free of specific genetic disorders, thereby enhancing the success rate and safety of in-vitro fertilization procedures[40].

Furthermore, certain medications might be more effective or pose more risks depending on an individual's genetic profile. A notable example is Warfarin, a commonly used drug for blood thinning. Precise dosing is crucial with Warfarin: Even a slight overdose can lead to bleeding, while too low a dose may not effectively prevent thrombosis. However, genetic variations in the *VKORC1* and *CYP2C9* genes can affect the efficacy and duration of action of warfarin (see fig. 1.1) and therefore require a dosage adjustment tailored to the patient's specific genetic variants[87].

Another area where genetic factors significantly impact treatment outcomes is cancer therapy. Influenced by genetic predisposition and environmental factors such as exposure to carcinogens (e.g., chemicals, radiation), inflammation, and viral infections, cancer arises from genetic and epigenetic alterations occurring in cells after birth, and each cancer tumor has its own unique genetic profile that leads to its uncontrolled growth [57]. Consequently, the success or failure of cancer treatment strongly depends on the genetic profile of the tumor. For instance, while some types of tumors can be treated successfully with drugs targeting the epidermal growth factor receptor (*EGFR*) gene,

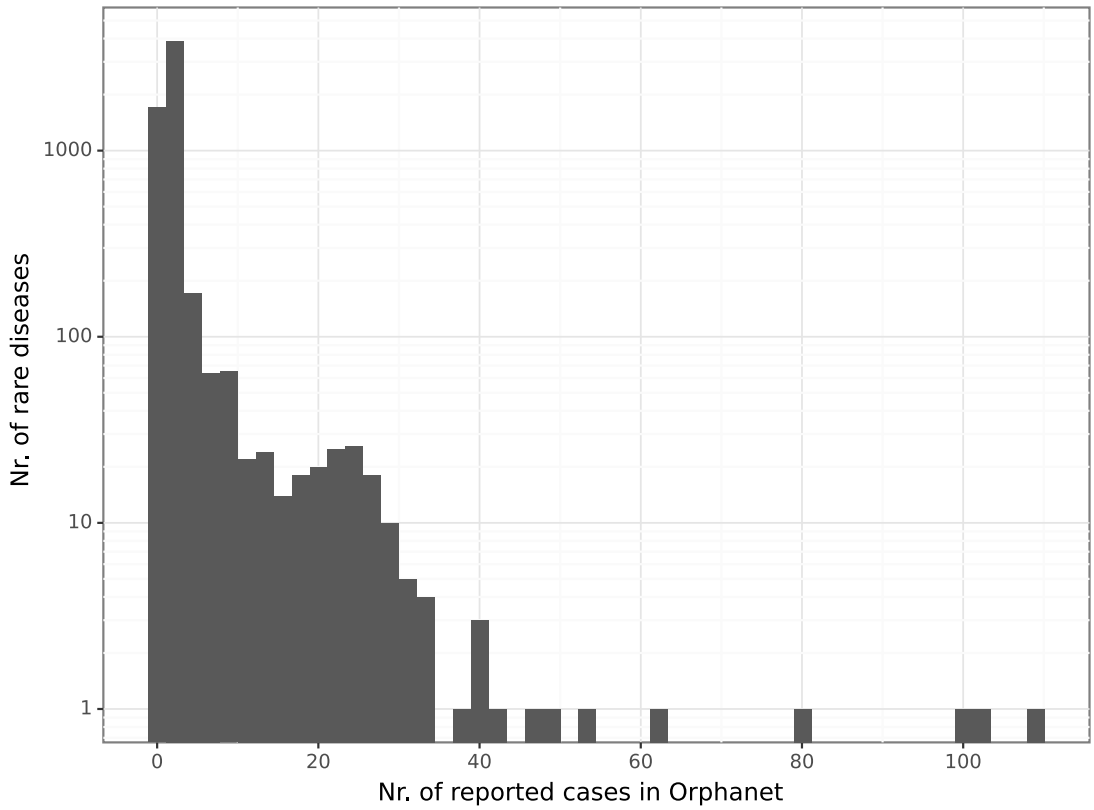

Figure 1.2: Distribution of reported cases among rare diseases in the Orphanet knowledge base. The Orphanet knowledge base is a comprehensive, international database that provides extensive information on rare diseases and orphan drugs[65].

mutations in another gene, *KRAS*, are known to confer resistance to these[152].

Understanding the genetic basis of diseases can also facilitate the development of new medications. Genes that are affected by disease-associated variants might play an important role in the disease and therefore can hold potential as targets for drug development[41].

Even more relevant is the identification of high-impact variants for the diagnosis and research of rare genetic disorders. Rare diseases are life-threatening or chronically debilitating conditions with a prevalence of less than 5 per 10,000 individuals[34].

Today we already know more than 10,000 rare diseases, with roughly 87% of them having a known or suspected genetic basis[85]. Despite their rarity, rare diseases collectively affect about 1 in 20 individuals, highlighting their significant impact on public health. However, given their low prevalence with many rare diseases having less than ten reported cases (fig. 1.2), rare diseases often receive less research attention and funding compared to more common conditions[67], resulting in only 5% of rare diseases having approved treatments as of today[38]. Identifying the disease-causing variants is not only

relevant for the correct diagnosis of patients suspected of having a rare genetic disease but also lays the foundation for understanding its molecular mechanisms and developing targeted therapies[134]. An example of a rare genetic disease is Duchenne muscular dystrophy, a severe condition characterized by the progressive weakening and tearing of muscle fibers that leads to premature death from cardiac and respiratory failure[28]. Duchenne muscular dystrophy is typically caused by pathogenic variants within the X-chromosomal gene *DMD*. *DMD* encodes a structural protein called dystrophin that prevents muscle fibers from tearing. Therefore, Duchenne muscular dystrophy can be diagnosed early by screening for pathogenic variants within the *DMD* gene[28]. Further, Duchenne muscular dystrophy is a so-called "recessive" disease, which means that one non-affected copy of the gene is sufficient to not get the disease.

Therefore, the disease-causal gene defect might be inherited from an asymptomatic mother[28]: Since women, unlike men, have two X chromosomes, the mother can carry a defective copy of *DMD* on one X chromosome and a functional copy on the other X chromosome. Identifying pathogenic variants in the mother's *DMD* gene would enable preventive measures to be taken in family planning, e.g. prenatal diagnostics for early detection or in-vitro fertilization to prevent the inheritance of the disease.

1.2 Most high-impact variants are rare

Seeking to identify variants related to heritable traits and diseases, large genome-wide association studies (GWAS) have identified thousands of common variants in the human genome that are associated with diseases[133, 148]. GWAS link single genetic variants to heritable traits by testing the association between the variant and the trait (fig. 1.3).

However, GWAS do not provide any information about the functional consequences of the associated variants and their causality. On the contrary, GWAS usually identify many co-occurring variants with the true causal variants, making it challenging to attribute the observed association to a specific variant. Further, most of the variants found by GWAS have only small effects on disease risk (fig. 1.4)[102]. Since selection removes variants with large pathogenicity from the population[75], high-impact variants are generally very rare in the population[131]. However, every human individual carries about 22,000 rare (i.e. less than 1 in 10,000 alleles) or private (i.e. unique to the individual) variants (fig. 1.5), and a single change at the wrong spot can cause a rare genetic disorder or increase the predisposition to a disease like Alzheimer or cancer by orders of magnitude[16]. GWAS require a large number of samples to find statistically significant associations between genetic variants and traits. This large sample size is needed to ensure that the results are not due to random chance. For rare variants, obtaining a sufficiently large cohort of individuals who carry these variants can be challenging. The rarer the variant, the more difficult it is to find enough individuals for a robust analysis. This limitation makes GWAS less suitable for studying extremely rare variants, as the statistical power to detect associations is significantly reduced. For novel variants that have never been

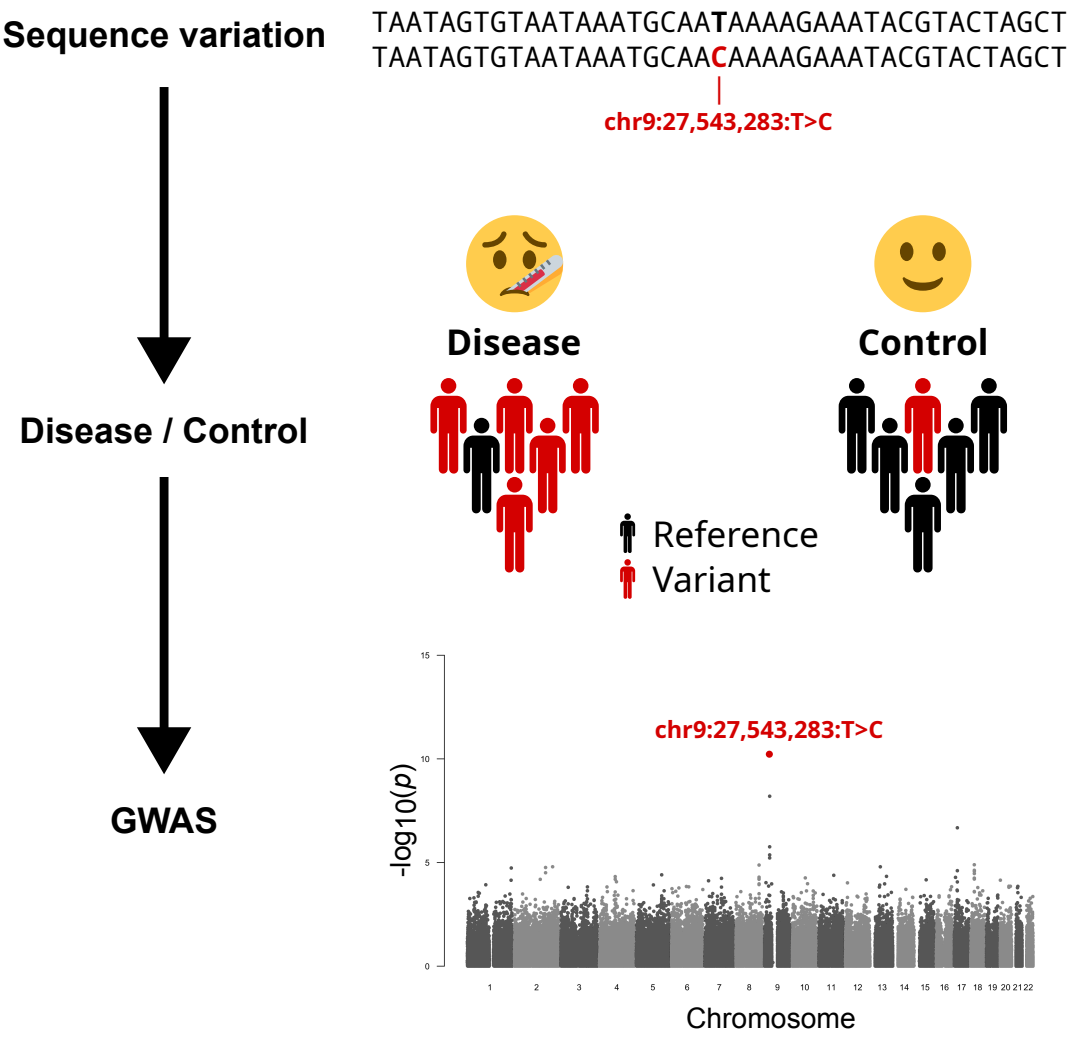

Figure 1.3: GWAS links single genetic variants to heritable traits. The figure shows an illustration of a genome-wide association study on a hypothetical disease. After genotyping a large cohort of individuals with and without the disease, GWAS perform a statistical test for each variant to assess its association with the disease, i.e. whether a certain genetic variant (red) is significantly more common in individuals with the disease compared to healthy control individuals[143]. GWAS can also test for association with quantitative traits, such as body height, by evaluating the correlation between genetic variants and trait measurements across the population.

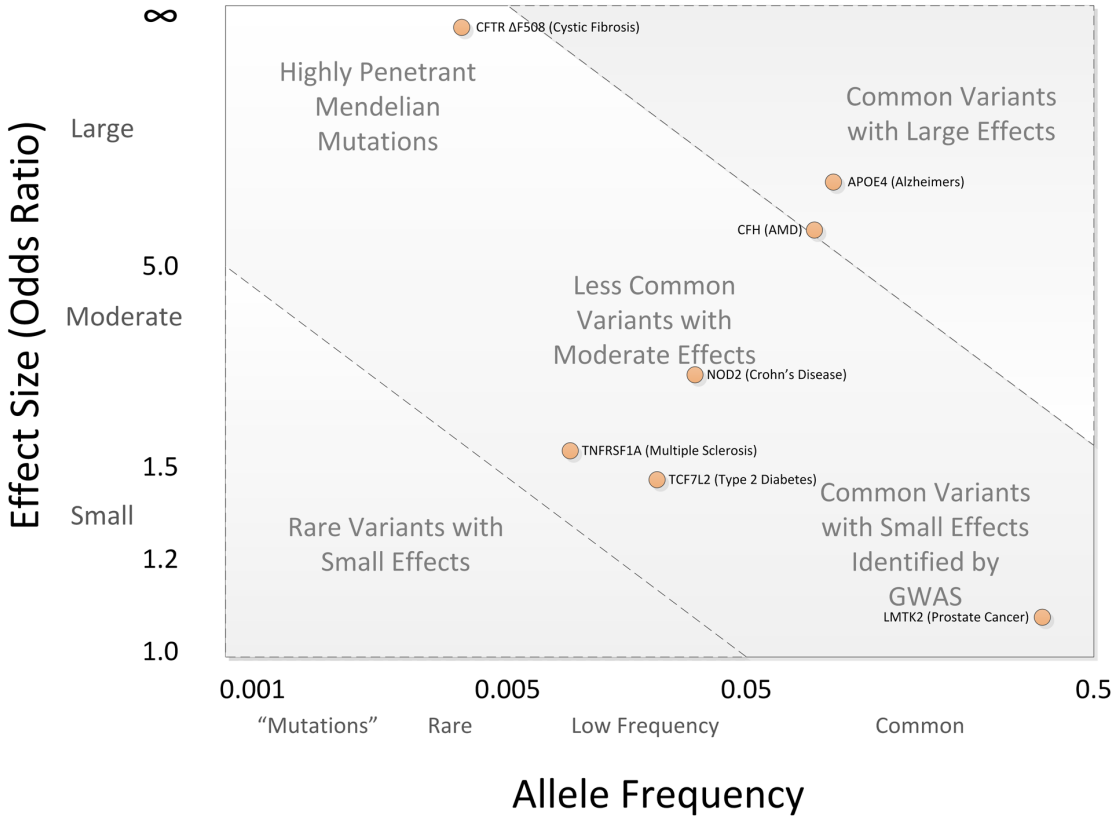

Figure 1.4: Variant effect size against variant allele frequency. High-impact variants causing Mendelian, i.e. genetic, disorders are usually rare (upper left), while most GWAS findings are associations of common variants with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines. Taken from [16].© 2012 Bush, Moore. Licensed under CC BY.

observed before in the population, conducting a GWAS is impossible.

Rare variant burden testing overcomes the limited statistical power of GWAS in detecting associations with rare genetic variants by aggregating rare variants within specific genomic regions, such as genes, into a gene-wise burden score and then testing the association between the burden score and some traits of interest[88]. This aggregation increases the overall frequency of the "variant signal" in the dataset, which can enhance the statistical power to detect an association, and reduces the multiple testing load. Rare variant burden testing depends on having many likely impactful variants with similar effects in the burden set (e.g. disrupting gene function which in turn increases the risk for disease)[88]. In the past, people assumed an inverse relationship between the minor allele frequency of the variant and its causality, i.e. the more rare a variant is, the higher impact it has. However, as stated before, every human individual carries about 22,000 rare or novel variants and most of these are benign[101]. Recent studies try to identify

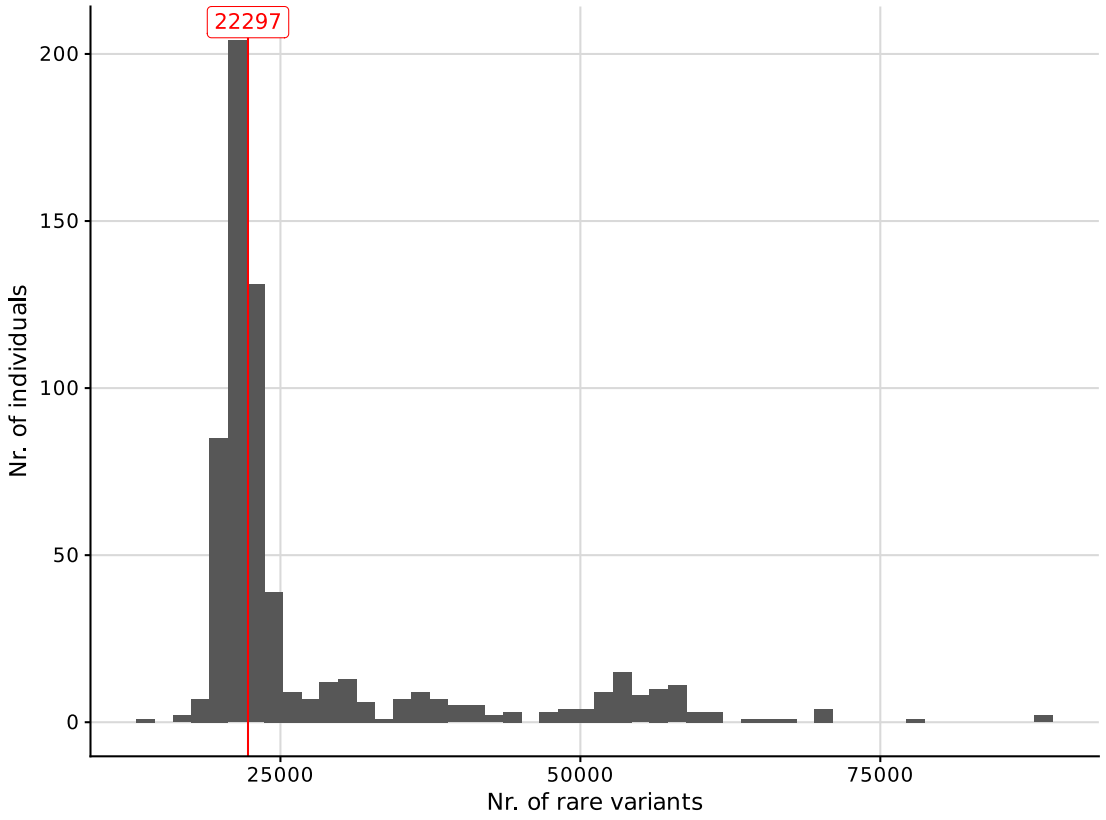

Figure 1.5: Distribution of rare variants among 635 individuals. The median number of rare variants per individual is 22,297 (red line). Numbers were calculated based on the GTEx dataset (see chapter 3).

potential high-impact variants and filter out low-impact variants by incorporating additional annotations that assess the functional and molecular impact of the variants[88, 106].

1.3 Biology of high-impact variants

Genetic variants can affect genes mainly in two ways: Either they influence the functionality of the gene's product or they disturb its regulation.

1.3.1 Coding variants

The function of a protein-coding gene is to produce messenger RNA (mRNA) that encodes the amino acid sequence of a protein (see section 2.2). If a variant changes the encoded amino acid sequence, the resulting protein might not fulfill its purpose anymore. The functional effect of protein-truncating variants is fairly well understood. Variants like stop-gains, frameshifts (small insertions or deletions causing a shift in the coding frame) or splice-site disruptions usually cause the truncation of the protein sequence by introducing premature stop codons and therefore a putative loss of function[70]. Human cells recognize premature stop codons during protein translation in the mRNA and degrade these faulty mRNA molecules in the nonsense-mediated decay (NMD) pathway (see section 2.3). For missense variants (variants that result in a different amino acid sequence without changing its length) it is more difficult to predict whether the resulting protein is still functional, but recent advances in protein structure prediction (AlphaMissense[19]) and sophisticated statistical modeling of protein sequence conservation (Evolutionary model of Variant Effect, EVE[46]; PrimateAI-3D[49]) have led to improved deleteriousness scores of missense variants.

1.3.2 Regulatory variants

However, protein-coding regions only cover 2% of the human genome. The vast majority of the human genome consists of non-coding DNA, which does not code for proteins but contains important regulatory regions that control when, where, and how much ribonucleic acid (RNA) is expressed from a gene. The regulation of gene expression happens at several stages, many of which can be influenced by genetic variation.

The first step in gene expression is making the DNA region accessible. DNA in eukaryotic cells is wrapped around histone proteins, forming a tightly packed structure called chromatin (fig. 1.6). For a gene to be expressed, the chromatin structure must be loosened or "opened" to allow access to the DNA sequence. Epigenetic modifications to the histone proteins and the DNA itself can switch chromatin between open and closed states, i.e. change which parts of the DNA are accessible[79]. The landscape of accessible DNA varies between different cell types and is a key factor in determining which genes are expressed in a particular cell[140].

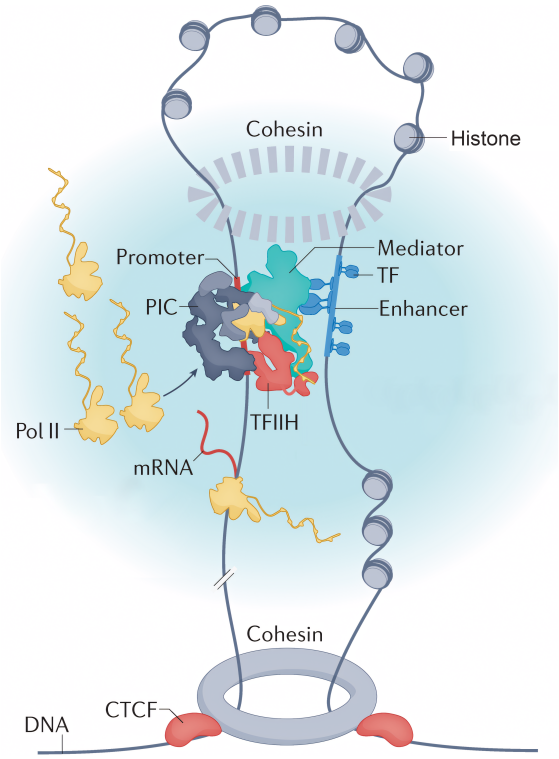

Figure 1.6: An enhancer-promoter interaction within a topologically associating domain formed by CTCF and cohesin. RNA polymerase II (Pol II) is positioned on the transcription start site of the promoter by the transcription preinitiation complex (PIC), which is formed by various general transcription factors such as TFIIH. The preinitiation complex gets stabilized by a mediator that is bound to the preinitiation complex and one or more transcription factors (TF) occupying the enhancer. Taken with permission from [123]. © 2022, Springer Nature Limited.

The next step is the recruitment of RNA polymerase, the enzyme responsible for transcribing DNA into RNA, in a transcription preinitiation complex. This recruitment is regulated by specific DNA sequences known as promoters, located near the start of genes. The preinitiation complex is formed by various promoter-binding general transcription factors that position the RNA polymerase on the transcription start site (TSS) of the gene. The RNA polymerase then moves along the DNA strand and transcribes its sequence into precursors of mRNA[56]. Genetic variants in the promoter sequence may affect the binding of transcription factors and therefore the rate of gene expression. RNA polymerase continues the transcription of DNA to mRNA until it reaches a polyadenylation signal, a specific sequence that then prompts the addition of a poly(A) tail to the mRNA transcript, eventually leading to the termination of transcription[141].

The preinitiation complex formation can be enhanced by other DNA elements called enhancers or suppressed by elements known as silencers. In the DNA sequence, enhancers and silencers can be located far from the gene they influence. The effect of enhancers and silencers on a gene depends on their spatial proximity to the gene's promoter, which is influenced by the formation of chromatin loops, so-called topologically associated domains (see fig. 1.6)[123]. Enhancers stabilize the preinitiation complex through a mediator that is bound to the preinitiation complex and one or more activating transcription factors occupying the enhancer, thereby increasing the gene expression. Repressive transcription factors that bind silencers can prevent enhancers from interacting with their target promoter, or they can interact directly with the promoter, thereby repressing gene expression[115]. Transcription factors play an important role in gene expression regulation. While general transcription factors are present in all cell types, activating and repressing transcription factors are often expressed in a cell type dependent manner and repress or activate gene expression accordingly[83]. Genetic variants in the transcription factor binding sites of enhancers can therefore lead to cell type dependent effects on gene expression.

Further, genes often have multiple alternative transcription start and termination sites that, together with alternative splicing, lead to the generation of tissue-dependent transcript isoforms[122]. Genetic variants that affect promoters, splice sites, and polyadenylation signals can therefore cause tissue-dependent changes in isoform expression based on the tissue specificity of these sites.

A final regulatory point of mRNA expression is the stability and degradation rate of mRNA molecules[147]. The rate at which mRNA molecules degrade is influenced by the binding of proteins and noncoding RNAs to certain sequences or structures within the mRNA. These proteins and noncoding RNAs recruit specific enzyme complexes that catalyze the degradation of the mRNA. Genetic variations can influence the stability of mRNA by affecting its structure and binding sites for these RNA-binding proteins and RNAs[90].

1.3.3 Integrating genomics with transcriptomics to identify high-impact variants

The effect of genetic variants, particularly on tissue and cell type specific gene expression, are often not straightforward to determine. Understanding the specific impact of a variant can be challenging, and in many cases, remains unknown or not fully understood[76, 109, 151].

Recently, statistical methods to call expression outliers from RNA sequencing (RNA-seq) data[14, 89, 124, 126] applied to large cohorts and various types of human tissues have enabled investigating the functional impact of variants on gene expression. By measuring aberrantly over- or underexpressed genes, i.e. gene expression that deviates strongly from normal patterns, and identifying rare genetic variants within these aberrantly regulated genes, algorithms can prioritize which variants may be the genetic cause of these expression outliers[39, 93]. Rare expression outlier associated variants identified by these algorithms have further been shown to be predictive of strong effects on phenotypic traits[132]. However, the requirement of sequencing the transcriptome in all relevant tissues of an affected individual can be challenging. Besides the additional costs for RNA-seq, diagnostics is often limited to clinically accessible tissues such as skin or body fluids as obtaining samples from other tissues like the brain or lung is significantly more invasive. Further, this approach does not generalize to unseen variants, as evaluating the impact of a genetic variant requires sequencing the transcriptome of an individual who carries it.

A key finding of integrative genomics and transcriptomics analysis is that certain types of rare genetic variants are highly enriched in gene expression outliers (fig. 1.7)[39, 93, 149]. Splice, frameshift, and stop variants often lead to nonsense-mediated decay of affected isoforms and are strongly enriched in underexpression outliers. Non-coding promoter variants can affect gene expression in both directions and are enriched in both over- and underexpression outliers. Among the structural variants, deletions and translocations are more enriched in underexpression outliers, while duplications and copy number variations are more common in overexpression outliers. These enrichment patterns are conserved and can be found in different cohorts, such as in the GTEx (Genotype-Tissue Expression) project[39, 93] and a mitochondrial disease study[149]. This raises the question of whether gene expression outliers might be predictable solely from the DNA sequence.

A method to predict aberrant expression in multiple tissues using only DNA sequence as input and generalizing to unseen variants could improve our ability to identify genetic variants with a high impact. This would in turn aid in the identification of disease-associated genes and pinpointing disease-causal variants in large genomic cohorts such as the UK Biobank (see section 2.12.4).

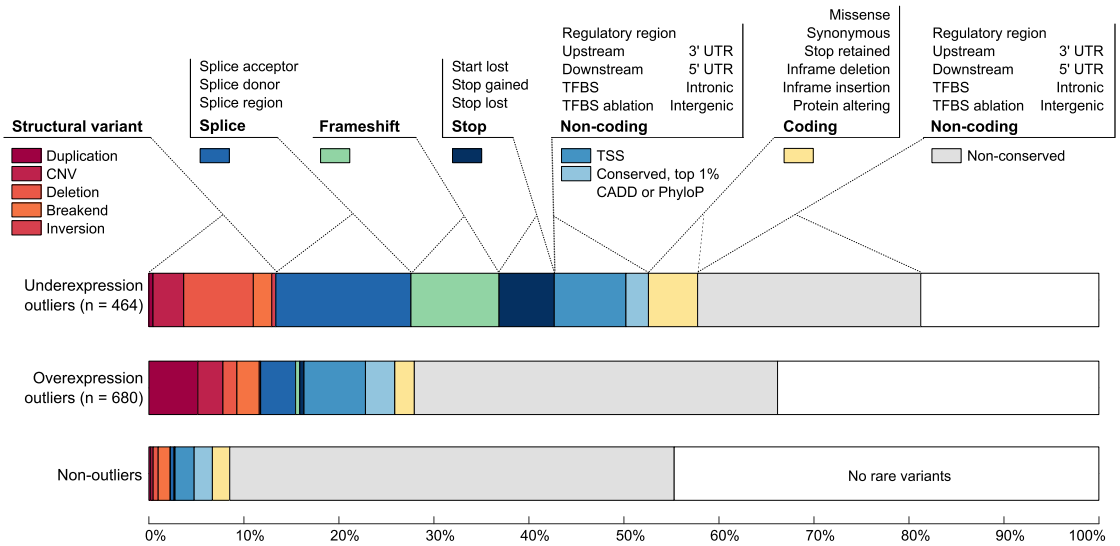

Figure 1.7: Proportion of gene expression outliers potentially explained by different classes of rare variants in a cohort of 449 individuals and 44 tissues from GTEx v6p. About half of the underexpression outliers can be explained by structural variants, non-coding variants between 250 bp upstream to 750 bp downstream of the TSS that affect for example transcription factor binding sites (TFBS), and variants associated with nonsense-mediated decay (splice, frameshift, stop). Among the overexpression outliers, copy number variations (CNV) and other structural variants, splice variants, and non-coding variants near the TSS can explain about a quarter. Breakend: translocation. Figure from [92].

1.4 Aims and scope of this thesis

The overall aim of this thesis is to enhance the identification of high-impact variants by assessing their effect on aberrant gene expression across human tissues and evaluate whether these changes in gene expression affect human traits. The contribution of my work is three-fold:

(i) Creating the first benchmark for aberrant gene expression across 48 human tissues and testing the predictive value of existing variant annotation tools, (ii) the development of AbExp, a method to predict aberrantly underexpressed genes in 48 human tissues from DNA sequence, and (iii) improving rare variant association studies and phenotype prediction with AbExp.

1.4.1 Benchmarking aberrant gene expression prediction in human tissues

Before this work there was no algorithm predicting aberrant gene expression based on an individual's genome. Also, the performance of existing variant annotation tools in the identification of gene expression outliers was not evaluated. To address this unmet need, I developed the first benchmark for predicting aberrantly expressed protein-coding genes by processing 11,096 RNA-seq samples with paired whole-genome sequencing data from 633 individuals across 48 tissues from GTEx. Then I used this benchmark to evaluate the performance of variant annotation tools attempting to solve similar tasks on predicting underexpression outliers.

1.4.2 AbExp: Predicting aberrant gene underexpression across human tissues

Given the lack of a specialized tool for aberrant expression prediction based on genetic variants, I next developed AbExp, a non-linear model combining various variant and tissue annotations to significantly improve the prediction of aberrant underexpression outliers across 48 human tissues. I applied AbExp on independent datasets to assess the generalizability of the predictions and the proficiency in distinguishing pathogenic from benign variants. Finally, I combined AbExp scores with gene expression measurements from clinically accessible tissues to predict aberrant expression in other tissues with improved precision.

1.4.3 Improving rare variant association testing and phenotype prediction with AbExp

Last, I demonstrate the application of AbExp in rare variant association testing and phenotype prediction of 40 blood traits in more than 200,000 individuals of the UK Biobank. Based on whole-genome sequencing data, AbExp scores offered supplementary information beyond the state-of-the-art putative loss of function classifier LOFTEE, improving

the discovery of significant gene-trait associations as well as significantly enhancing phenotype prediction.

Most of the work presented in this thesis is published as a preprint[61].

2 Background

2.1 Genetic variants

2.1.1 Types of genetic variants

Genetic variants can be categorised according to their length and their mechanism of origin[81, 111]. Single nucleotide variants (SNVs), which represent changes to individual base pairs in the reference sequence, and small base insertion or deletions (INDELs), i.e. variants that change the length of the reference sequence, affect one to 50 base pairs in a single event (fig. 2.1). Longer events are referred to as structural variants (SVs). Copy number variations (CNVs) such as long deletions (DEL) and duplications (DUP) as well as insertions (INS) are called imbalanced SVs as they change the length of the genome. In contrast, inversions (INV) and translocations (TRA) are balanced SVs as they retain the length of the genome.

2.1.2 Zygosity of variants

Variants can be present in two forms with respect to an individual's paired chromosomes: If the variation (allele) is present on only one of the two homologous chromosomes, the individual is heterozygous for that genetic locus ("heterozygous variant"). If the same variation (allele) is present on both homologous chromosomes at a particular locus, the individual is homozygous for that genetic locus ("homozygous variant")[3].

2.2 Central dogma of biology

The central dogma of biology describes the flow of genetic information from deoxyribonucleic acid (DNA) over RNA to protein (fig. 2.2)[3]. Within DNA, specific segments are called genes. Each gene provides instructions for creating a functional product, for example, a protein.

The process begins with transcription, during which a gene's DNA sequence is copied to create a pre-mature messenger RNA (pre-mRNA). Transcription is initiated by the binding of RNA polymerase. Beginning at the transcription start site (TSS), it reads the template DNA strand in the 3' to 5' direction (fig. 2.2, red strand), while synthesizing a complementary strand in the 5' to 3' direction (fig. 2.2, blue strand), translating adenine to uracile (U), thymine to adenine (A), cytosine to guanine (G), and guanine to cytosine (C). As the last step in transcription, a 5' cap and a poly-A tail are being added.

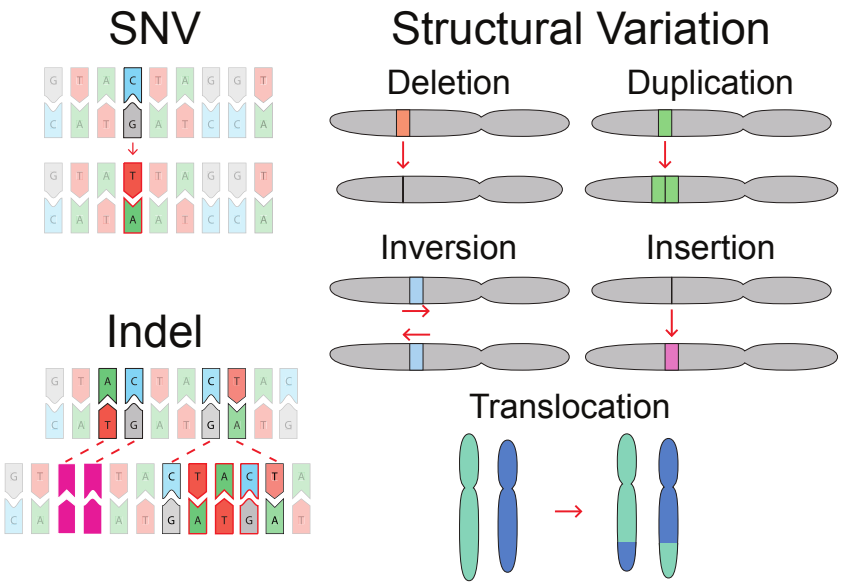

Figure 2.1: Overview of types of genetic variation. The top chromosome denotes the reference chromosome, while the variation is highlighted and shown below. Single nucleotide variants (SNV) and small insertions and deletions (INDEL) represent changes that affect one to 50 base pairs in a single occurrence. Longer events are referred to as structural variants (SV). These events include deletions (DEL), duplications (DUP), inversions (INV), insertions (INS), translocations (TRA), and complex combinations of these basic variant types. Figure from [111].

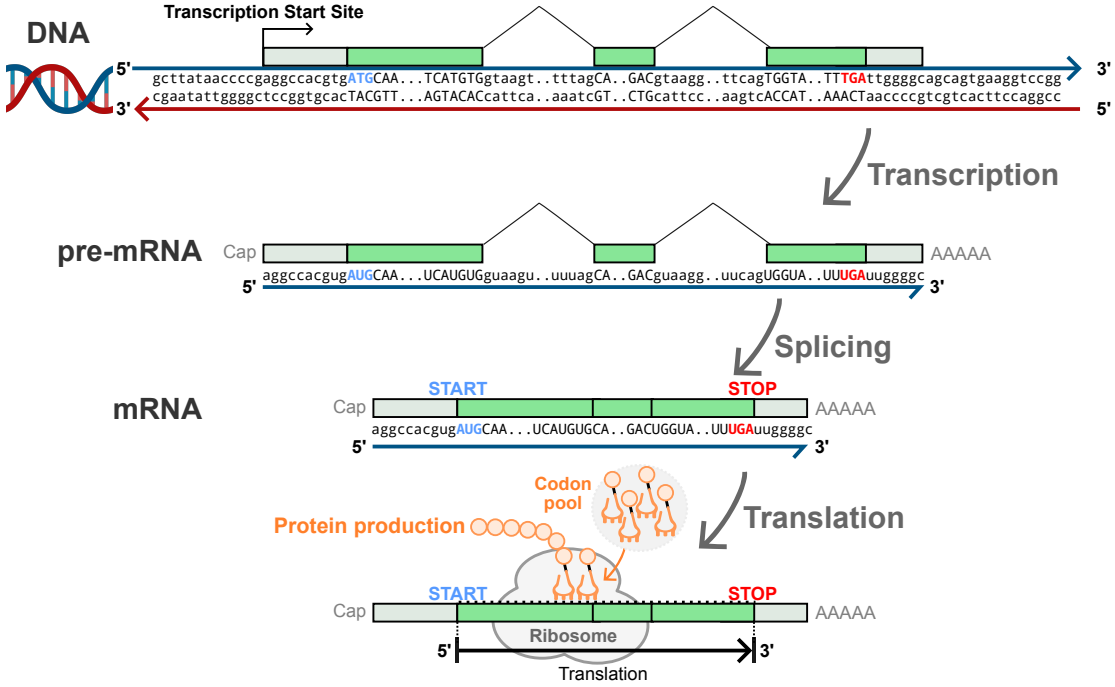

Figure 2.2: The central dogma of biology. DNA first gets transcribed into premature messenger RNA (pre-mRNA), beginning at the transcription start site. Splicing removes introns from the pre-mRNA, leaving a mature mRNA molecule from which proteins can be translated. Finally, a ribosome binds and translates, beginning at an AUG start codon, the coding sequence into a protein sequence until it reaches one of several stop codons, e.g. UGA.

Next, splicing removes introns and joins exons to form mature messenger RNA (mRNA). Finally, ribosomes translate the mRNA sequence into a protein sequence. The ribosome moves along the mRNA molecule in steps of three nucleotides called codons, beginning at a start codon (AUG). Each codon corresponds to a specific amino acid, start, or stop signal (fig. 2.3). Notably, the mapping of codons to amino acids is redundant, i.e. multiple codons can encode the same amino acid. As each codon is exposed, a complementary transfer ribonucleic acid (tRNA) molecule carrying the corresponding amino acid enters the ribosome and pairs with the mRNA codon through base pairing. The amino acid carried by the tRNA is added to the growing polypeptide chain through peptide bond formation. When the ribosome encounters a UAA, UAG, or UGA codon, translation terminates.

2.3 Nonsense-mediated decay of mRNA

Nonsense-mediated decay (NMD) is a quality-control mechanism in eukaryotic cells that serves to degrade mRNA molecules containing premature termination codons (PTCs)[3]. PTCs are stop codons that occur prematurely within the coding region of mRNA, leading

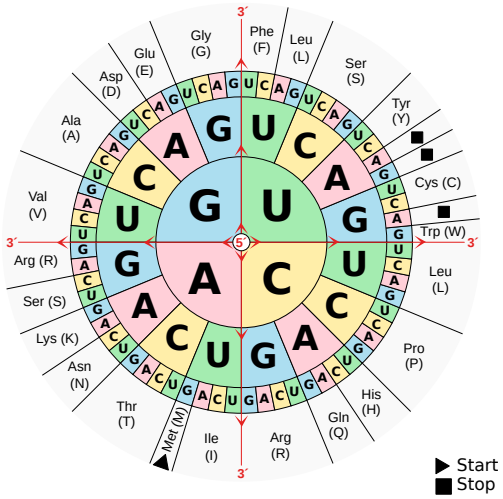

Figure 2.3: The standard RNA codon table organized in a wheel. The sequence is read from 5' to 3' direction, e.g. AUG → Methionine. 64 possible codon combinations encode 22 different amino acids as well as the start and stop codon. Figure from [108] (public domain).

to the production of truncated or incomplete proteins. PTCs frequently result from frameshifts, splice site disruptions, or missense mutations. NMD acts as a quality control mechanism to prevent the translation of abnormal or potentially harmful truncated proteins.

The process of NMD involves several steps. After splicing, exon junction complexes (EJC) remain at the splice sites to mark successful splicing. When the first ribosome starts translating the mature mRNA into a protein sequence, it strips off these EJCs. If the ribosome reaches a premature termination codon (STOP) before stripping off the last EJC, nonsense-mediated decay is induced by UPF proteins that recognize the EJC, and the mRNA gets decayed (fig. 2.4).

Not all PTCs induce NMD[94, 95]. The NMD machinery typically fails to recognize PTCs located more than 50 nucleotides upstream of the last exon-exon junction ("50bp rule"). Additionally, PTCs in the last exon of a transcript do not activate NMD due to the lack of exon junction complexes ("last exon rule"). For the same reason, also transcripts with a single exon do not induce NMD. Furthermore, NMD is inhibited in very long exons, generally exceeding approximately 400 nucleotides ("long exon rule"). Moreover, PTCs positioned within 150 nucleotides from the start codon often do not elicit NMD, likely due to translation re-initiation ("start-proximal rule"). Figure 2.5 shows an illustration of these rules.

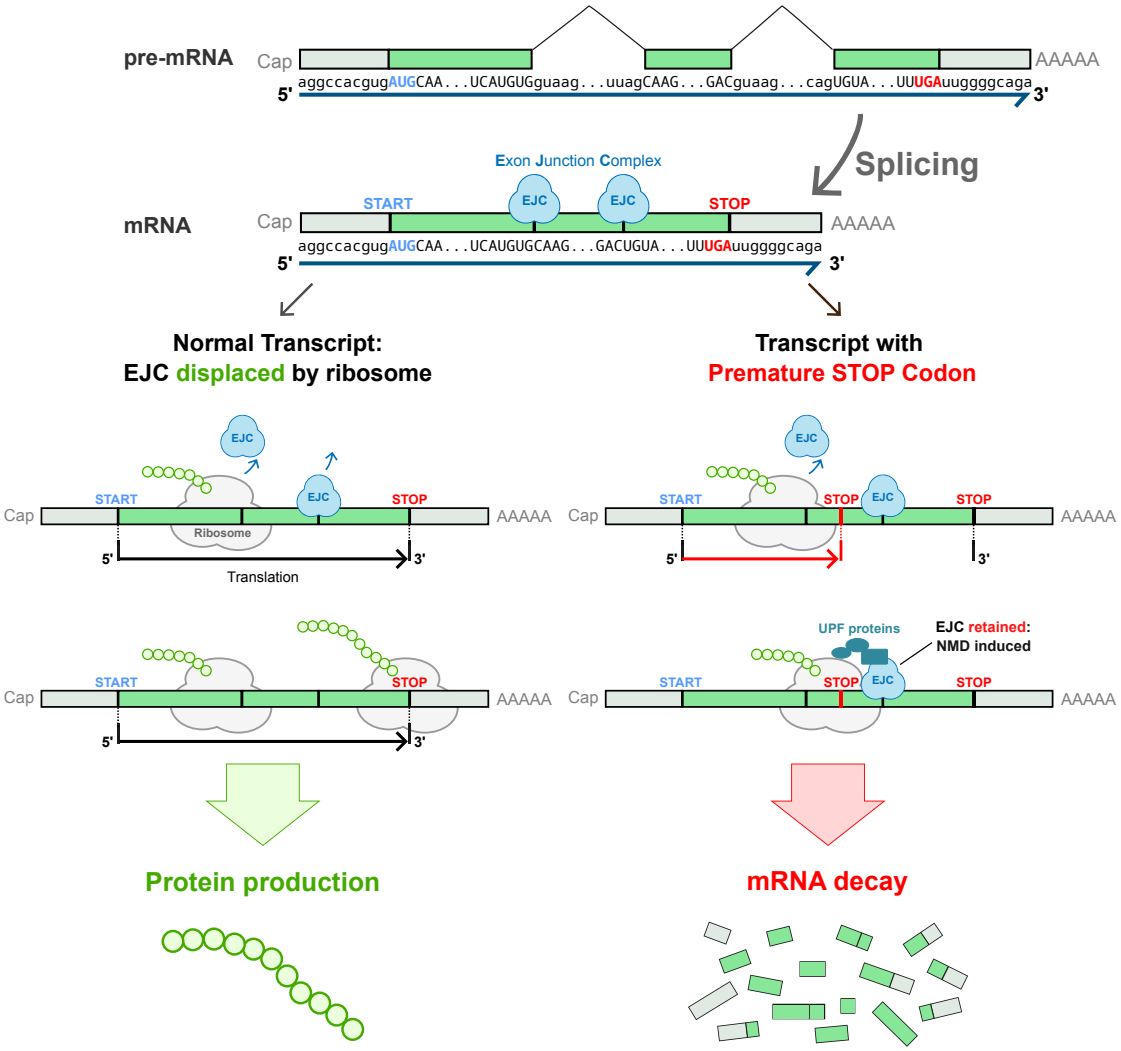

Figure 2.4: Overview of nonsense-mediated decay pathway. After splicing, exon junction complexes (EJC) remain at the splice sites to mark successful splicing. During normal translation, these complexes get stripped off by the ribosome (left side). If the ribosome reaches a premature termination codon (STOP) before stripping off the last EJC, nonsense-mediated decay is induced by UPF proteins that recognize the EJC, and the mRNA gets decayed (right side).

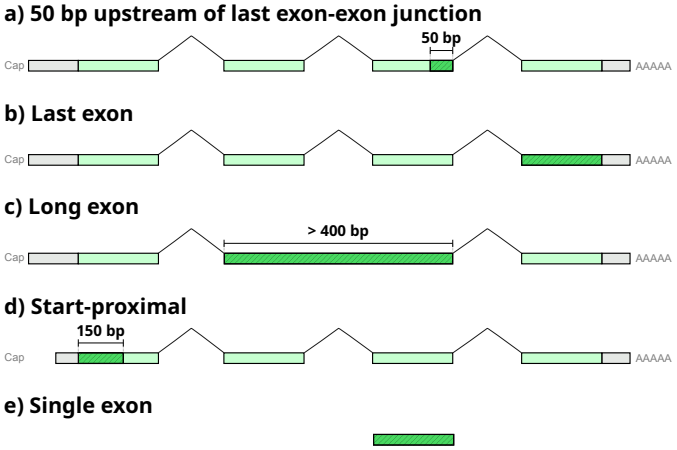

Figure 2.5: Premature termination codons (PTCs) in certain regions of the gene do not induce nonsense-mediated decay (NMD). Transcripts harboring PTCs within the highlighted regions typically escape NMD. **(a)** 50 bp rule, **(b)** last exon rule, **(c)** long exon rule, **(d)** start-proximal rule, **(e)** single exon.

2.4 Next-Generation Sequencing

Next-generation sequencing (NGS), also referred to as short-read sequencing[4], is a technology for rapid and cost-effective sequencing of DNA and ribonucleic acid (RNA) by simultaneously sequencing millions of DNA fragments in parallel[130]. The sequencing process begins by extracting DNA from the sample of interest. The DNA is then fragmented into smaller pieces using various methods, such as sonication or enzymatic digestion (fig. 2.6a). These fragments serve as the templates for sequencing.

Next, sequencing adapters are ligated to the ends of the fragmented DNA molecules (fig. 2.6b). Adapters contain sequences that are complementary to those on the sequencing platform and enable the DNA fragments to bind to the surface of the sequencing platform. Adapters also contain sequences necessary for subsequent steps, such as priming sites for PCR amplification, a barcode that identifies the fragment, and other implementation-dependent elements.

Then, the DNA fragments with adapters attached are amplified using polymerase chain reaction (PCR) and immobilized on a solid surface, such as a glass slide or a flow cell (fig. 2.6c). This step increases the amount of DNA available for sequencing and generates clusters of identical DNA fragments. PCR amplification is crucial for ensuring that there is enough DNA for detection during sequencing and for improving the signal-to-noise ratio.

Finally, the immobilized fragments are sequenced in an iterative cycle of nucleotide incorporation and imaging (fig. 2.6d). During each cycle, fluorescently labeled nucleotides are added to the DNA fragments, and the incorporation of each nucleotide is detected by imaging. The fluorescent signal is recorded and used to determine the sequence of the

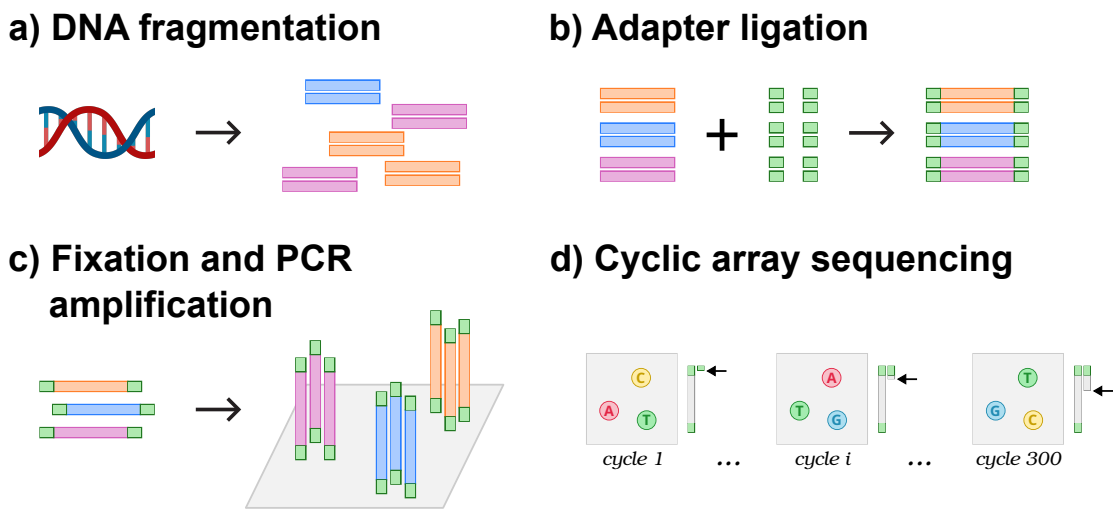

Figure 2.6: Next-generation sequencing. **(a)** First, the DNA is fragmented into smaller pieces. **(b)** Next, the fragments get adapter sequences attached that typically contain priming sites, fragment bar codes, and other implementation-dependent elements. **(c)** The fragments get fixed to a surface (e.g. solid substrate or beads) and amplified by PCR. **(d)** Clusters of immobilized fragments are sequenced in an iterative cycle of nucleotide incorporation and imaging, sequencing typically between 75-300 nucleotides of the fragment ends.

DNA fragment. This cycle of nucleotide incorporation, imaging, and washing is repeated multiple times to sequence millions of DNA fragments simultaneously, typically reaching sequence lengths of about 300 bases.

These sequences read from the DNA fragment ends are called “reads”. DNA fragments are typically sequenced from both ends by sequencing both complementary strands. This allows the identification of pairs of reads from the same DNA fragment based on the fragment barcode embedded in the adapter (“paired-end reads”). Paired-end sequencing provides additional information about the DNA fragment, such as its length [63].

2.4.1 DNA-seq

RNA sequencing (DNA-seq) is used to analyze the genome, either specific regions like exonic regions (Whole-Exome Sequencing) or all regions (Whole-Genome Sequencing). After next-generation sequencing (NGS), the resulting fragment reads are typically aligned to some reference genome (fig. 2.7a). While de novo assembly of the sequences is possible, it requires deep sequencing of the sample to gain enough coverage (i.e. number of times a specific nucleotide in a DNA or RNA sequence is read, also called “read depth”). It is less costly and faster to align against some shared high-quality reference genome and also allows for easy detection of short sequence variants. For humans, there are multiple versions of reference genomes, such as (sorted by age) GRCh37, GRCh38, and T2T-CHM13[1]. Further, special tools can detect large structural variations within RNA

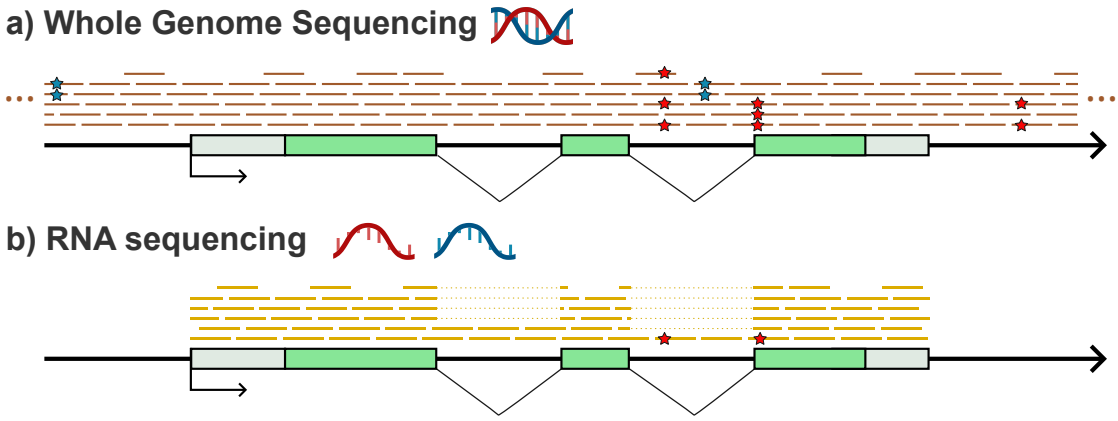

Figure 2.7: Read sequence alignment to the reference genome. In both figures, stars denote variants where the measured sequence differs from the reference sequence. The color of the star denotes the strand. **(a)** Alignment of whole genome DNA sequencing reads. DNA-seq reads cover nearly the whole genome. **(b)** Alignment of RNA sequencing reads. RNA sequencing (RNA-seq) reads cover transcribed regions, and reads from spliced RNA can also skip intronic regions ("split-read", dotted lines). RNA-seq data aligns to transcribed regions of the genome, providing information about RNA abundance, alternative splicing, and sequence variation within these regions.

sequencing (DNA-seq) data[55] or detect variant phasing, i.e. variants that affect the same strand[21].

2.4.2 RNA-seq

RNA sequencing (RNA-seq) is used to analyse the transcriptome by sequencing complementary DNA (cDNA) synthesized from RNA[63]. RNA-seq reads cover only transcribed regions such as exons and introns (fig. 2.7b), providing information about gene expression levels (read coverage within the gene), alternative splicing events (split reads), and sequence variants within the transcribed regions.

The number of fragments measured from a gene depends not only on the expression level but also on its length and the sequencing depth. Longer genes will have more fragments sequenced than shorter genes. Similarly, when comparing the number of fragments between experiments, the number of measured fragments per gene depends on the total amount of sequenced fragments. Normalization techniques such as TPM (Transcripts Per Million) or FPKM (Fragments Per Kilobase of transcript per Million mapped reads) help mitigating these biases when comparing gene expression levels across

samples or experiments[153]:

$$FPKM_i = \frac{q_i}{\frac{l_i}{10^3} \cdot \frac{\sum_j q_j}{10^6}} = \frac{q_i}{l_i \cdot \sum_j q_j} \cdot 10^9 \tag{2.1}$$

$$TPM_i = \frac{q_i / l_i}{\sum_j \left(q_j / l_j\right)} \cdot 10^6 \tag{2.2}$$

Here, i is a feature (e.g. gene or transcript), q_i is the number of raw fragment counts of the feature, and l_i is the length of the feature.
The FPKM measure can easily be converted to TPM:

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j}\right) \cdot 10^6 \tag{2.3}$$

2.5 OUTRIDER: Aberrant gene expression calling

OUTRIDER (Outlier in RNA-Seq Finder) is a statistical method to detect gene expression outliers in RNA-seq datasets[14]. OUTRIDER uses a denoising autoencoder to model RNA-seq fragment count expectations with a negative binomial distribution (fig. 2.8). Once the expected counts are established, OUTRIDER identifies aberrantly expressed genes whose observed read counts significantly deviate from these expectations.

Specifically, OUTRIDER models the probability of the observed fragment count $x_{s,g}$ for every gene g in a sample s as:

$$P(x_{s,g}|\mu_{s,g}, \theta_{t(s),g}) = \mathrm{NB}(x_{s,g}|\mu_{s,g}, \theta_{t(s),g}) \tag{2.4}$$

where:

- $\mu_{s,g}$ is the expected fragment count
- $\theta_{t(s),g}$ is the dispersion parameter for the gene g in the tissue of sample s $t(s)$

OUTRIDER further outputs:

- the biological coefficient of variation:

$$\mathrm{BCV}_{t(s),g} = \frac{1}{\sqrt{\theta_{t(s),g}}} \tag{2.5}$$

- the $\log_2$-transformed fold-change of the observed fragment count compared to the expected fragment count:

$$\log_2 \mathrm{FC} = \log_2(x) - \log_2(\mu) \tag{2.6}$$

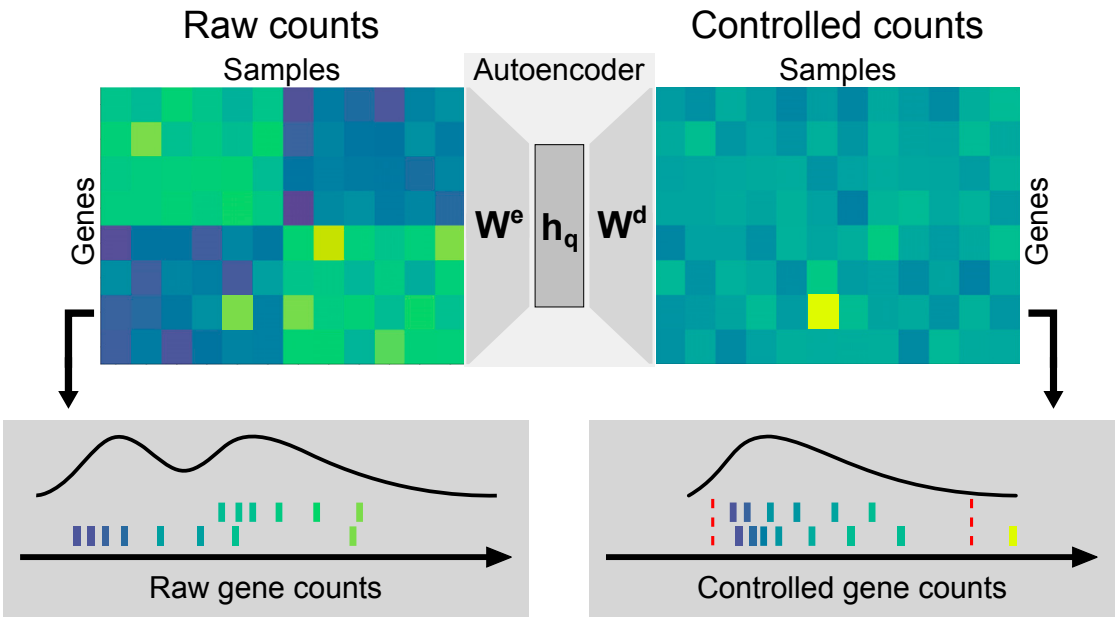

Figure 2.8: Context-dependent outlier detection with OUTRIDER. *"The algorithm identifies gene expression outliers whose read counts are significantly aberrant given the covariations typically observed across genes in an RNA-seq dataset. This is illustrated by a read count (left panel, fifth column, 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."* Taken with permission from [14]. © 2018 American Society of Human Genetics.

- the nominal p-value
- the False Discovery Rate using the Benjamini-Yekutieli method[11]

2.6 Precision-Recall Curve and Average Precision

A precision-recall curve is a graphical representation used in machine learning and information retrieval to evaluate the performance of a classification model, particularly in binary classification tasks. It illustrates the trade-off between precision and recall at different thresholds for a given classifier.

Precision (or Positive Predictive Value) describes the fraction of true positive predictions among the total number of positive predictions (true positives + false positives). Recall (or Sensitivity or True Positive Rate) describes the fraction of true positive predictions among the actual number of positive cases (true positives + false negatives). By ranking the classifier predictions according to their confidence scores or probabilities, one can compute the precision and recall at different thresholds. The precision-recall curve illustrates this trade-off.

To summarize the overall performance of the classifier across all possible thresholds, one can calculate the area under the precision-recall curve (AUPRC), also known as the ap[154]:

$$P = \frac{T_P}{T_P + F_P} \tag{2.7}$$

$$R = \frac{T_P}{T_P + F_N} \tag{2.8}$$

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{2.9}$$

Here, T_P is the number of true positive predictions, F_N is the number of false negative predictions, P is the precision and R the recall. The average precision metric provides a single numerical value that quantifies the classifier's ability to balance precision and recall across all threshold levels.

2.7 Supervised learning

Supervised learning is a fundamental concept in machine learning where the objective is to develop a predictive model that maps input features to output labels based on example input-output pairs ("training data"), aiming to accurately predict the target labels for new, unseen data.

Let us assume we would like to predict N observed values Y based on a matrix of N

observed values of D independent predictor variables X:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N \tag{2.10}$$

$$X = \begin{pmatrix} x_{1,1} & \dots & x_{1,D} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \dots & x_{N,D} \end{pmatrix} = (x_1, \dots, x_D) \in \mathbb{R}^{N \times D} \tag{2.11}$$

The goal of supervised learning is to develop a predictor which accurately models the relation between X and Y, such that it can be applied to predict target labels for new, unseen data.

2.7.1 Linear regression

Linear regression models the relationship between Y and X using a linear equation[60]:

$$Y = \beta_0 + \sum_{d=1}^{D} \beta_d x_d + \varepsilon \tag{2.12}$$

Here, $\beta \in \mathbb{R}^D$ represents a vector of coefficients, and $\varepsilon \in \mathbb{R}^N$ represents a vector of normally distributed errors. fig. 2.9 shows an illustration of a linear model with one predictor variable x_1.

The goal of linear regression is to estimate the coefficient vector $\hat{\beta}$ that minimizes the difference between the observed values of Y and the values predicted by the linear model $\hat{Y} = \beta_0 + \sum_{p=1}^{D} \beta_p X_p$. This is often done using the method of least squares, which minimizes the residual sum of the squared differences between the observed values $y_i \in Y$ and and predicted values $\hat{y}_i \in \hat{Y}$:

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \mathrm{RSS}(\beta) \\ &= \arg\min_{\beta} \left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \right) \\ &= \arg\min_{\beta} \left(\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{D} \beta_j x_{ij})^2 \right) \end{aligned} \tag{2.13}$$

Assuming that X has full column rank, i.e. that the columns of X are linearly independent, it is possible to obtain an analytic solution for $\hat{beta}$:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.14}$$

This equation is also known as the ordinary least squares (OLS) estimator.

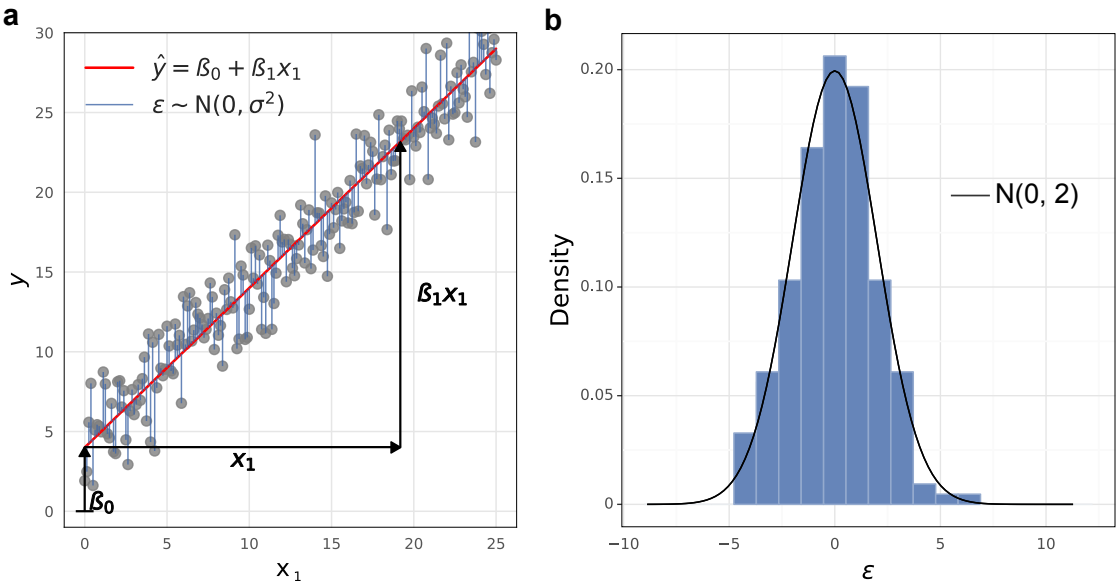

Figure 2.9: Illustration of a linear regression model with one predictor variable. y, observed values; $\hat{y}$, model prediction; β_0, intercept; β_1, model weights; ε, normally distributed error between y and $\hat{y}$. **(a)** Scatter plot of y against x with a linear equation shown as red curve and errors ε highlighted as blue vertical lines. **(b)** Density histogram of errors ε. The black curve shows the probability density function of a normal distribution with a mean at 0 and a standard deviation of 2.

2.7.2 Logistic regression

Logistic regression is a variant of linear models used for binary classification tasks, where the observed values Y are categorical and have only two possible outcomes, typically represented as 0 and 1[12].

In logistic regression, the goal is to model the probability that a given observation belongs to a particular category. The logistic function, also known as the sigmoid function σ, is used to map the linear combination of the independent variables to the range (0, 1). The logistic function is defined as:

$$\begin{aligned} P(Y = 1|X) &= \sigma\left(\beta_0 + \sum_{d=1}^{D} \beta_d x_d\right) \\ &= \frac{1}{1 + e^{-(\beta_0 + \sum_{d=1}^{D} \beta_d x_d)}} \end{aligned} \tag{2.15}$$

, where $P(Y = 1|X)$ is the probability that the observed values Y equal 1 given the values of the independent predictor variables X.

The loss function typically used in logistic regression is the categorical cross-entropy loss function. Given a binary classification problem where the true class labels are $y_i \in \{0, 1\}$ and the predicted probabilities of class 1 are $P(y_i = 1|X_i) = \hat{y}_i$, the aim is to minimize the categorical cross-entropy loss function with respect to β:

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \mathrm{CCE}(\beta) \\ &= \arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)\right) \end{aligned} \tag{2.16}$$

The logistic loss function penalizes the model more heavily when the predicted probability deviates from the true label:

- If $y_i = 1$, the loss function penalizes the model more when $\hat{y}_i$ is close to 0.
- If $y_i = 0$, the loss function penalizes the model more when $\hat{y}_i$ is close to 1.

Unlike ordinary linear regression, there is no analytical solution for $\hat{beta}$. Instead, the loss function needs to be minimized with iterative approaches such as gradient descent or iterative reweighted least squares[12].

2.7.3 Elastic Net Regularization

Elastic Net regularization is a technique used in regression analysis and machine learning to prevent overfitting and improve the generalization performance of the model. It combines the penalties of both Lasso (L1 regularization) and Ridge (L2 regularization) regression methods[155].

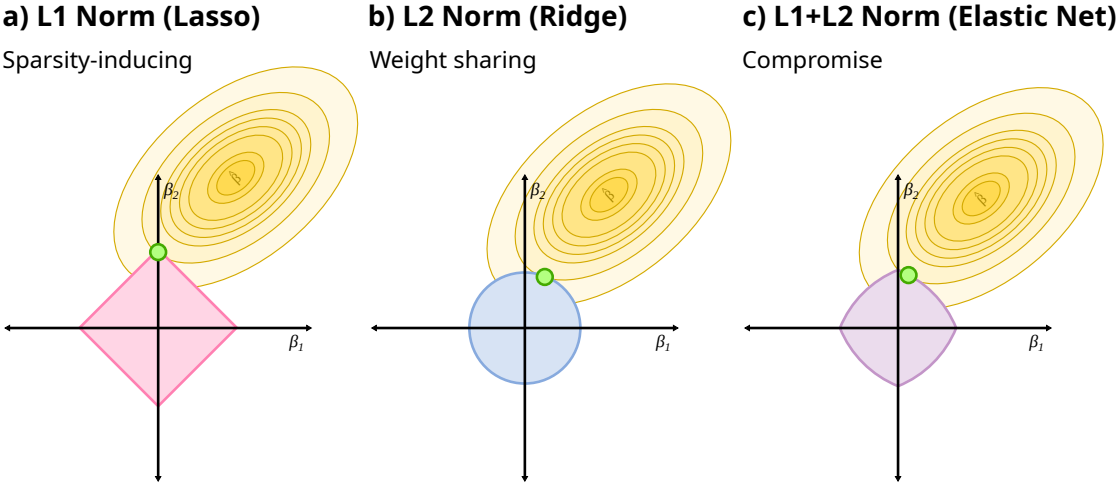

Figure 2.10: Contour plot of least squares cost function and various regularization methods for a two-dimensional model. The orange elliptical contours show the loss for settings of a two-dimensional model in terms of its parameters β_1 and β_2, with $\hat{\beta}$ denoting the optimal solution without regularization. The red **(a)**, blue **(b),** and purple **(c)** regions show the various penalty terms. By introducing regularization, the optimal solution is being constrained towards the origin (all zero β vector). Increasing λ strengthens the regularization, leading to smaller penalty term regions. The green point represents the optimal regularized parameters, found at the geometric minimal intersection between the penalty region and the parameter contour. In contrast to L2 regularization, solutions obtained through L1 regularization often tend to "slide to the corners", aligning with the axes and causing subsets of the β vector to become exactly zero.

Elastic Net regularization can be added to an arbitrary cost function by incorporating penalty terms for both the L1 (Lasso) and L2 (Ridge) norms of the model parameters. Let us assume we want to regularize some cost function $L(\beta)$. The general form of the cost function with Elastic Net regularization $J(\beta)$ can be expressed as:

$$J(\beta) = L(\beta) + \lambda_1 \sum_{d=1}^{D} |\beta_d| + \lambda_2 \sum_{d=1}^{D} \beta_d^2 \tag{2.17}$$

Here, $\sum_{d=1}^{D} |\beta_d|$ denotes the L1 norm or Lasso penalty term, which penalizes the absolute values of the coefficients. $\sum_{d=1}^{D} \beta_d^2$ denotes the L2 norm or Ridge penalty term, which penalizes the squared values of the coefficients. λ_1 and λ_2 control the strength of the Lasso and Ridge penalties, respectively.

Lasso regression tends to produce sparse models by driving some coefficients to exactly zero, effectively performing feature selection. Ridge regression tends to shrink the coefficients towards zero without eliminating them entirely, which helps reduce the impact of multicollinearity. Elastic Net regularization combines the advantages of Lasso and Ridge regression (see also fig. 2.10).

2.7.4 LightGBM: Gradient boosting decision trees

Light Gradient Boosting Machine (LightGBM) is an open-source gradient boosting framework that uses tree-based learning algorithms to solve supervised learning tasks, particularly classification and regression problems, with a focus on fast and efficient training[74].

Decision tree methods construct a hierarchical tree structure where each internal node represents a decision based on the value of a specific feature or attribute, and each leaf node represents the predicted outcome or class label[60]. Internal nodes of the tree represent decisions that recursively partition the feature space by splitting it along the axes of the input variables. This partitioning process continues until a stopping criterion is met. Leaf nodes of the tree represent the outcome or class label within each region created by the splits. To make predictions, decision tree methods assign a constant value to each leaf node. For regression problems, this constant value is typically the mean or median of the target variable within the region. For classification problems, it may be the most frequent class label. Decision trees are popular due to their simplicity, interpretability, and ability to handle non-linear relationships and interactions between variables.

Gradient boosting is a non-linear machine learning technique that sequentially adds weak learners to an ensemble to produce a strong learner, where each weak learner corrects its predecessor[48]. The term "gradient" in gradient boosting refers to the optimization technique used to minimize the loss function by iteratively fitting new models to the residuals of the previous models.

Training an ensemble model of gradient-boosted decision trees begins with initializing the ensemble with a simple model, usually a single decision tree or a constant value representing the target variable's initial approximation. Gradient boosting then builds the ensemble model sequentially by adding new decision trees to correct the errors made by the existing ensemble. Each new model is trained on the residuals (the differences between the predicted values and the actual target values) of the current ensemble. Figure 2.11 shows an example of this gradient boosting procedure.

For regression problems, LightGBM uses by default the Mean Squared Error (MSE) between N observed values $y_i \in Y$ and predicted values $\hat{y}_i \in \hat{Y}$ as loss function:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.18}$$

Gradient-boosting decision trees are non-linear models that can theoretically approximate any function by adding a large enough amount of weak learners to the ensemble. This involves the risk of overfitting by adding too many or too complex decision trees to the ensemble. LightGBM has several configurable parameters that control the complexity of the model, the regularization applied, and the training process to prevent overfitting and improve generalization:

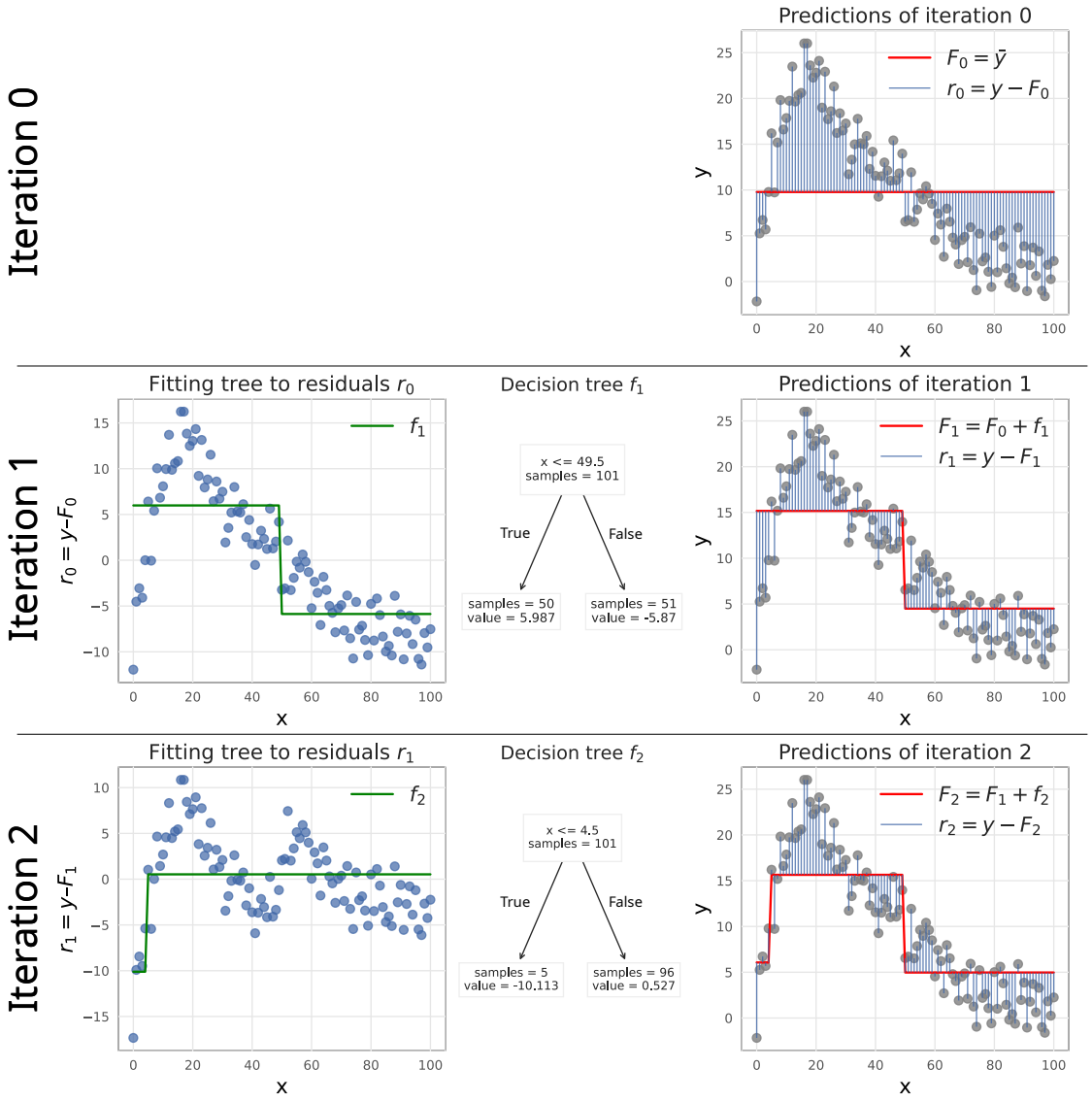

Figure 2.11: Building an ensemble model of decision trees with gradient boosting. The aim is to create an ensemble model F which predicts a target variable y from an input variable x. The ensemble is initialized (F_0) with a first estimator, which can be an initial weak learner or the average of the prediction target y. Then, based on the residual errors in the previous iteration (r_{i-1}), additional weak estimators (f_i) are trained and added to the ensemble model, iteratively refining its predictions.

- **num_leaves:** Maximum number of leaves in one tree. A higher value can make the model more complex and may lead to overfitting.
  Default: `31`.

- **n_estimators:** Number of boosting iterations (trees) to be run in the gradient boosting model.
  Default: `100`.

- **max_depth:** Maximum depth of the tree. A smaller value limits the complexity of the model and helps prevent overfitting. A value of -1 means there is no limit to the depth.
  Default: `-1`.

- **min_child_samples:** Minimum number of data points required in a child (leaf) node. If a split results in a child node with fewer samples than `min_child_samples`, the split is not considered.
  Default: `20`.

- **min_child_weight:** Minimum sum of instance weight (hessian) needed in a child (leaf) node.
  Default: `0.001`.

- **min_split_gain:** Minimum loss reduction required to make a split. If the gain is lower than `min_split_gain`, the split is not performed.
  Default: `0`.

- **reg_alpha:** L1 regularization term on weights. This adds a penalty to the absolute value of coefficients to prevent overfitting.
  Default: `0`.

- **reg_lambda:** It is the L2 regularization term on weights. This adds a penalty to the square of coefficients to prevent overfitting.
  Default: `0`.

2.8 Likelihood Ratio Test

A likelihood ratio test is a statistical hypothesis test used to compare the goodness of fit of two competing statistical models, typically nested within each other. The likelihood ratio test assesses whether adding additional parameters to a simpler model significantly improves its fit to the data[35].

Let us assume, we would like to know whether some outcome variable Y is linearly dependent on a predictor variable $x_i \in X$ (section 2.7.1). Therefore, our null hypothesis H_0 is that there is no linear relationship between the outcome variable Y and the predictor variable x_i, i.e. the coefficient of x_i in the linear regression model is zero ($\beta_i = 0$).

Correspondingly, our alternative hypothesis H_a is that there is a linear relationship, i.e. $\beta_i \neq 0$. In other words, we want to test whether a "full" model with estimated parameters $\hat{\beta}$ is fitting Y significantly better than a "reduced" model with parameters $\hat{\beta}_0$ excluding one or more of predictor variables:

$$H_0 : \left|\hat{\beta} - \hat{\beta}_0\right| = 0 \tag{2.19}$$

$$H_a : \left|\hat{\beta} - \hat{\beta}_0\right| > 0 \tag{2.20}$$

To conduct this hypothesis test, we can use the likelihood ratio test (LRT) as our statistical method. The likelihood ratio test compares the likelihood of the data under the full model $L_0(\beta)$ to the likelihood under the reduced model $L_a(\beta)$.
In the context of linear regression, the likelihood function represents the probability of observing the given set of observed values Y given the predictor values X and the parameters of the model β as the product of the individual error probabilities[12]:

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{N} N(\epsilon_i | \mu = 0, \sigma^2) \\ &= \prod_{i=1}^{N} N(y_i - \hat{y}_i | \mu = 0, \sigma^2) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \sum_{d=1}^{D} \beta_d x_{id})^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-N/2} \exp\left(-\sum_{i=1}^{N} \frac{(y_i - \beta_0 - \sum_{d=1}^{D} \beta_d x_{id})^2}{2\sigma^2}\right) \end{aligned} \tag{2.21}$$

Here, $N(\epsilon_i | \mu = 0, \sigma^2)$ denotes the probability of the normal distributed error term, and σ the variance of the error term. Smaller errors will be closer to zero and thus are more likely than larger error terms. Figure 2.12 shows an illustration of this idea. Notably, maximizing the likelihood $L(\beta)$ is equivalent to minimizing the RSS (see also section 2.7.1).

The test statistic of the LRT is $-2\ln(\Lambda)$, where Λ denotes the likelihood ratio between H_0 and H_a:

$$\begin{aligned} -2\ln\Lambda &= -2\ln\frac{L_0}{L_a} \\ &= -2 \cdot [\ln(L_0) - \ln(L_a)] \\ &= \ \ 2 \cdot [\ln(L_a) - \ln(L_0)] \end{aligned} \tag{2.22}$$

H_a has to be at least as likely as H_0, therefore, the value range of $-2\ln(\Lambda)$ is in the range of $[0, +\infty)$. The higher the difference between H_0 and H_a, the more significant is the difference. According to the Wilks theorem[146], if H_0 holds true, this test statistic will be asymptotically χ^2-distributed if the number of observations approaches ∞, with

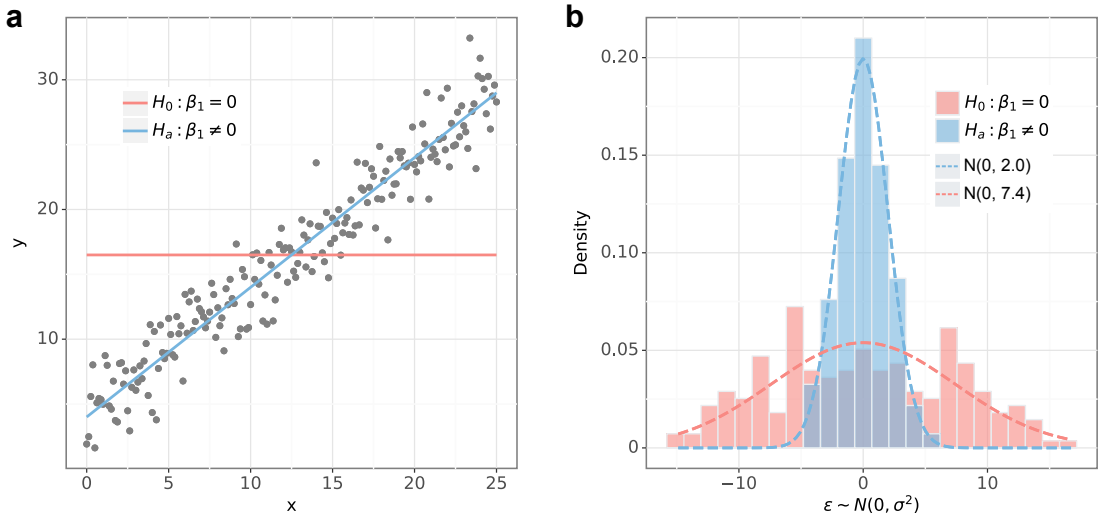

Figure 2.12: Example of two alternative models fitting some observed values y. (a) Observed values y against a predictor variable x, with $y = \beta_0 + \beta_1 x + \epsilon$. The red curve shows the null hypothesis of y being independent of x. The blue curve shows the alternative hypothesis of y being linearly dependent on x. **(b)** Density histogram of model errors ϵ for null and alternative hypothesis from (a). The dashed lines show a normal distribution fitted to the observed errors. The closer the error is to 0, the higher its probability.

a degree of freedom df equal to the number of additional parameters used by the full model:

$$\begin{aligned} -2\ln\Lambda &\sim \chi^2_{\Delta df}, \\ \Delta df &= df_{H_a} - df_{H_0} \end{aligned} \tag{2.23}$$

Therefore, a χ^2 test can be used to check if there is a significant difference between H_0 and H_a[116]. H_0 will be rejected with a significance level α if $(-2\ln\Lambda)$ is higher than the $(1-\alpha)$ quantile of the $\chi^2_{\Delta df}$ distribution:

$$\begin{aligned} \alpha &= P(X > -2\ln\Lambda \mid H_0) \\ &= \chi^2_{\Delta df}(X > -2\ln\Lambda) \\ &= 1 - \chi^2_{\Delta df}(X \leq -2\ln\Lambda) \\ &= 1 - \mathrm{CDF}_{\chi^2_{\Delta df}}(-2\ln\Lambda) \end{aligned} \tag{2.24}$$

$$\Rightarrow (1-\alpha) = \mathrm{CDF}_{\chi^2_{\Delta df}}(-2\ln\Lambda) \tag{2.25}$$

Here, $\mathrm{CDF}_{\chi^2_{\Delta df}}(x)$ denotes the cumulative distribution function of a χ^2 distribution with Δdf degrees of freedom.

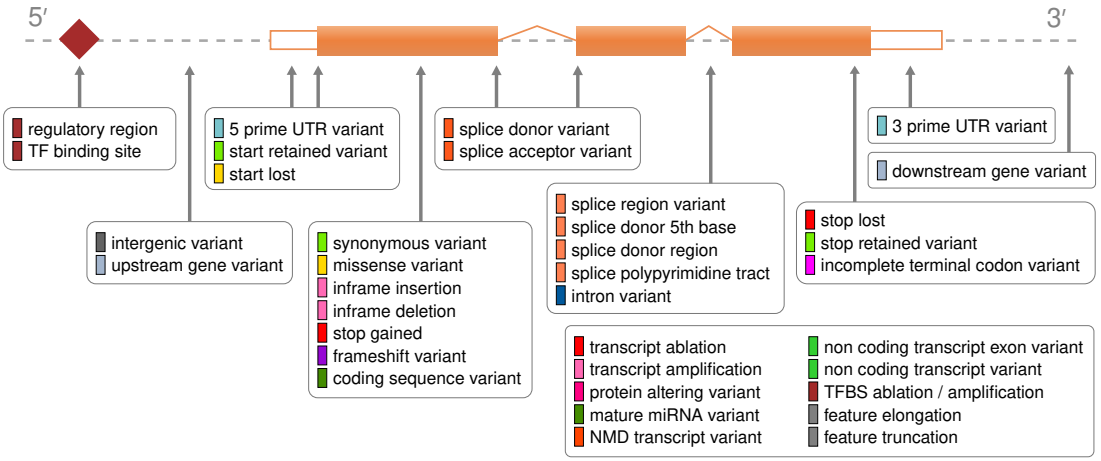

Figure 2.13: Variant consequences calculated by Ensembl VEP. Figure from [33]. © EMBL-European Bioinformatics Institute. Licensed under Apache 2.0.

2.9 Ensembl VEP

Ensembl Variant Effect Predictor (VEP) is a tool developed by the Ensembl project that predicts the functional consequences of genetic variants (mutations) in genomic data. It is widely used in genomics research and clinical genetics to interpret the impact of genetic variations on genes, transcripts, and proteins. VEP takes genetic variants, such as single nucleotide variants (SNVs), insertions, deletions, and structural variants, and annotates them with information about their potential effects on genes and gene products according to the Sequence Ontology[30]. This includes predicting whether a variant affects protein-coding regions (e.g., missense, nonsense, frameshift mutations), splice sites, regulatory elements, and other genomic features. Figure 2.13 shows an overview of the annotated consequences in Ensembl VEP version 108.

VEP can be extended with plugins to include additional variant annotations, such as LOFTEE or CADD. Notably, VEP offers a plugin to identify stop-gained variants that escape NMD, based on the rules outlined in section 2.3. In particular, the NMD plugin of VEP will assign the sequence ontology term “NMD_escaping_variant”, if a stop-gained variant is located on the last coding exon of a transcript, within the last 50 bp of the penultimate coding exon of a transcript, within the first 100 bp of the coding sequence of a transcript, or within an intronless transcript[32].

2.10 LOFTEE

The Loss-Of-Function Transcript Effect Estimator (LOFTEE) is a tool designed to predict a high-confidence subset of loss-of-function variants within protein-coding genes, particularly those likely to induce nonsense-mediated decay[70].

To distinguish between "high-confidence" (HC) and "low-confidence" (LC) variants, LOFTEE applies a series of filters. For stop-gained and frameshift variants, it considers factors such as proximity to transcript ends and splice site characteristics. Variants that are near the end of transcripts or land in exons with non-canonical splice sites may be flagged as "low-confidence". Similarly, for splice-site variants, LOFTEE filters out variants that only affect splicing of untranslated regions (UTRs) or are not predicted to affect a donor site. It also considers evolutionary aspects and intron characteristics to filter out certain variants. In addition to assessing known LoF-inducing variants, LOFTEE also makes predictions of other splice (OS) variants that may cause LoF by disrupting normal splicing patterns. It uses logistic regression models to evaluate whether variants significantly disrupt extended splice sites and SVM models to predict variants creating de novo donor splice sites leading to frameshifts.

2.11 Combined Annotation Dependent Depletion (CADD)

Combined Annotation Dependent Depletion (CADD) is a tool designed to estimate the deleteriousness of genetic variants in the human genome[77]. It integrates diverse annotations from genomic features and functional elements to assign a variant-specific score, which reflects the predicted impact of the variant on protein function, gene expression, and regulatory elements. It was trained on a binary distinction between simulated de novo variants and variants that have arisen and become fixed in human populations.

As input features, CADD utilizes annotations obtained using Ensembl VEP (section 2.9), conservation and selection scores, epigenetic information, and other annotations available for subsets of variants. Examples of annotations include transcript information like distance to exon-intron boundaries, DNase hypersensitivity, transcription factor binding, expression levels in commonly studied cell lines and amino acid substitution scores for protein coding sequences[120, 121]. Notably, CADD v1.6 introduced additional splicing annotations from SpliceAI[66] and AbSplice-DNA[144] to improve the improved the prediction of splicing effects[120].

2.12 Datasets

2.12.1 GTEx

The Genotype-Tissue Expression (GTEx) project is a large-scale genome and transcriptome sequencing study to investigate gene expression and regulation and its relationship to genetic variation across different human tissues[23, 53]. As part of the GTEx project, samples were therefore collected from various tissues obtained post-mortem from more than 800 healthy donors, followed by sequencing the genomes of the donors and the transcriptomes of the tissue samples. Figure 2.14 shows an overview of the tissue sampling sites used in the GTEx project.

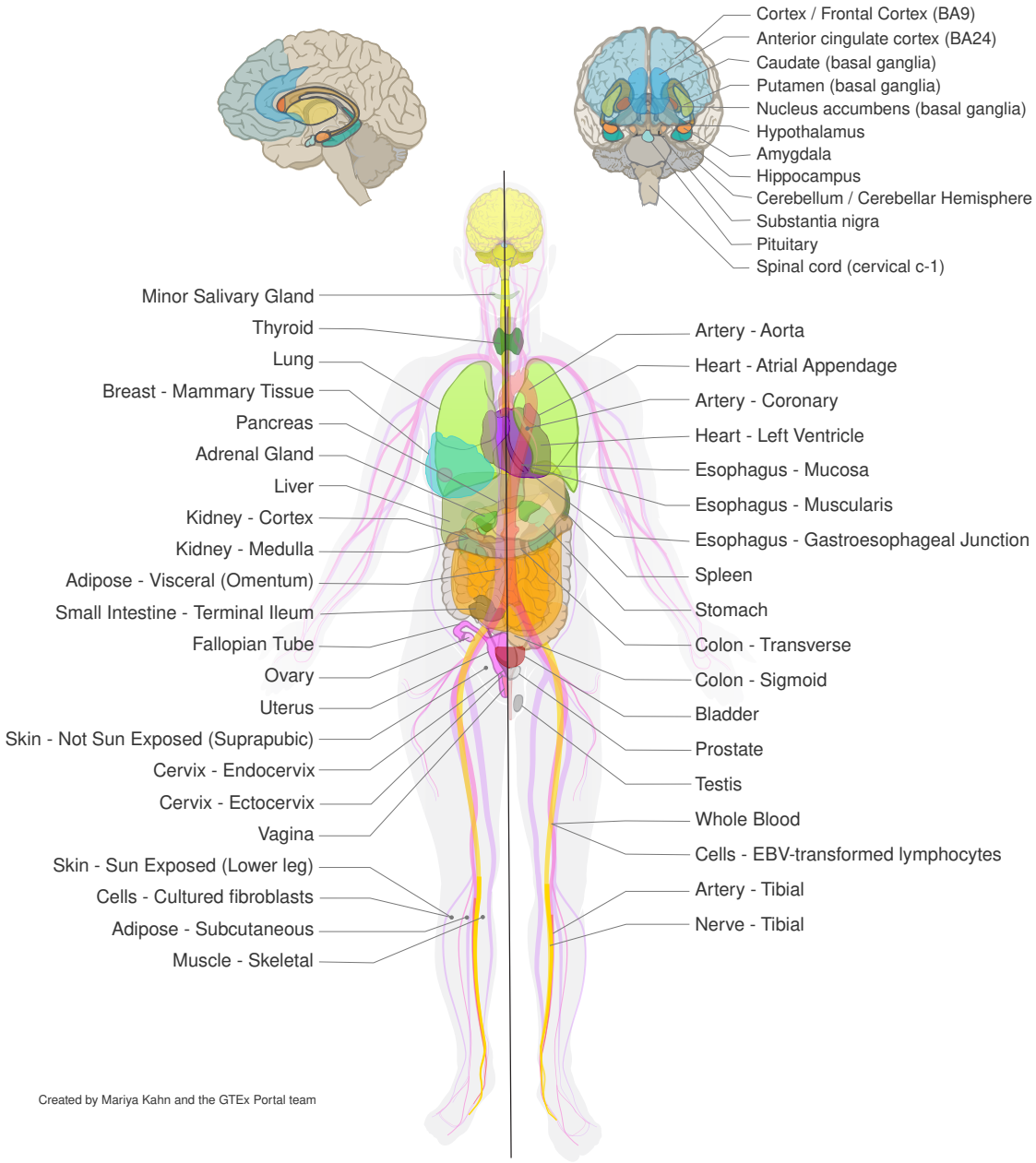

Figure 2.14: Tissue sampling sites used in the GTEx project. Obtained with permission from [54]. © 2021 Broad Institute of MIT and Harvard.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS and provides raw and processed data at https://gtexportal.org/. Note that raw sequencing data as well as genotypes are only available after application through dbGaP.

2.12.2 Answer ALS

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease causing the degeneration of motor neurons. Patients with ALS suffer from progressive muscle weakness which leads to paralysis and ultimately death from respiratory failure[58]. In Europe, ALS occurs with a frequency of 2-3 cases per 100,000 individuals. Although there are possible therapies to slow down disease progression, as of today, ALS disease cannot be cured and typically ends with death from respiratory failure after 24-48 months. A popular exception was the astrophysicist Stephen Hawking, who lived with ALS for 55 years after his initial diagnosis[118].

The Answer ALS Research Program is a comprehensive research initiative aimed at uncovering the biological mechanisms underlying ALS[10]. To this end, the Answer ALS program provides a biological and clinical resource of patient iPSC-derived motor neurons, and longitudinal clinical and smartphone data from over 1,000 patients with ALS. Induced pluripotent stem cells (iPSCs) are pluripotent stem cells that are generated directly from adult blood cells. These iPSCs can derive into any of the cell types that make up the body, such as motor neurons. In the Answer ALS program, patients provide a blood sample from which iPSC-derived motor neurons are generated. These iPSC-derived motor neurons are then analysed with a multi-omics approach including whole-genome sequencing, RNA transcriptomics, ATAC-sequencing and proteomics.

2.12.3 Mitochondrial disease dataset

Mitochondrial diseases are a group of genetic disorders that result from abnormalities in mitochondria[52]. Mitochondria are membrane-bound cell organelles responsible for producing ATP (adenosine triphosphate) through oxidative phosphorylation of glucose[3]. ATP is the central unit that powers chemical reactions within our cells. Symptoms of mitochondrial disorders can vary widely and may include muscle weakness, vision and hearing problems, developmental delays, neurological issues, and organ dysfunction. The severity and specific symptoms of mitochondrial disorders depend on which cells and organs are affected and to what extent mitochondrial function is impaired. Mitochondrial diseases can be caused by genetic variation in both nuclear and mitochondrial DNA. It is estimated that between 5 and 20 out of 100,000 individuals are affected by mitochondrial diseases[52].

The mitochondrial disease dataset used in this work is described in Yépez et al. (2022)[149] and contains whole-exome sequencing and skin fibroblasts transcriptome se-

quencing data of more than 300 individuals suspected to suffer from a mitochondrial disease.

2.12.4 UK Biobank

The UK Biobank is a large-scale biomedical database and research resource that includes genetic, clinical, and lifestyle information from around 500,000 participants in the United Kingdom[136]. It aims to improve the prevention, diagnosis, and treatment of various diseases by providing researchers with access to health records, questionnaires, and genome sequence data of the participants.

In this work, I used data from about 200,000 individuals for whom both whole-exome sequencing and microarray genotyping data were available at the time of the study[138]. Note that, as of today, there are whole-exome and additionally whole-genome sequencing data for more than 490,000 participants available in the UK Biobank[91].

2.12.5 ClinVar

ClinVar is a freely accessible public archive of reports on the relationships among human variations and phenotypes, with supporting evidence[86]. It aggregates information about genomic variation and its relationship to human health, including interpretations of the clinical significance of variants. ClinVar provides classifications for the clinical significance of genetic variants, ranging from pathogenic (causing or contributing to disease), likely pathogenic, benign (not associated with disease), likely benign, uncertain significance, and others.

2.12.6 GnomAD

The Genome Aggregation Database (gnomAD) is a comprehensive resource that aggregates and harmonizes exome and genome sequencing data from a wide range of large-scale sequencing projects[70]. It contains data from both population-based and disease-focused studies, providing information on genetic variants observed in diverse human populations. gnomAD is widely used by researchers and clinicians to assess the frequency and distribution of genetic variants across different populations and to identify rare variants associated with diseases. Figure 2.15 shows an overview of the number of whole-exome and whole-genome sequencing samples included in different GnomAD releases as well as its predecessor, ExAC. Note that ExAC and the GnomAD v2 release are mapped to the GRCh37 reference sequence while newer releases are mapped to GRCh38.

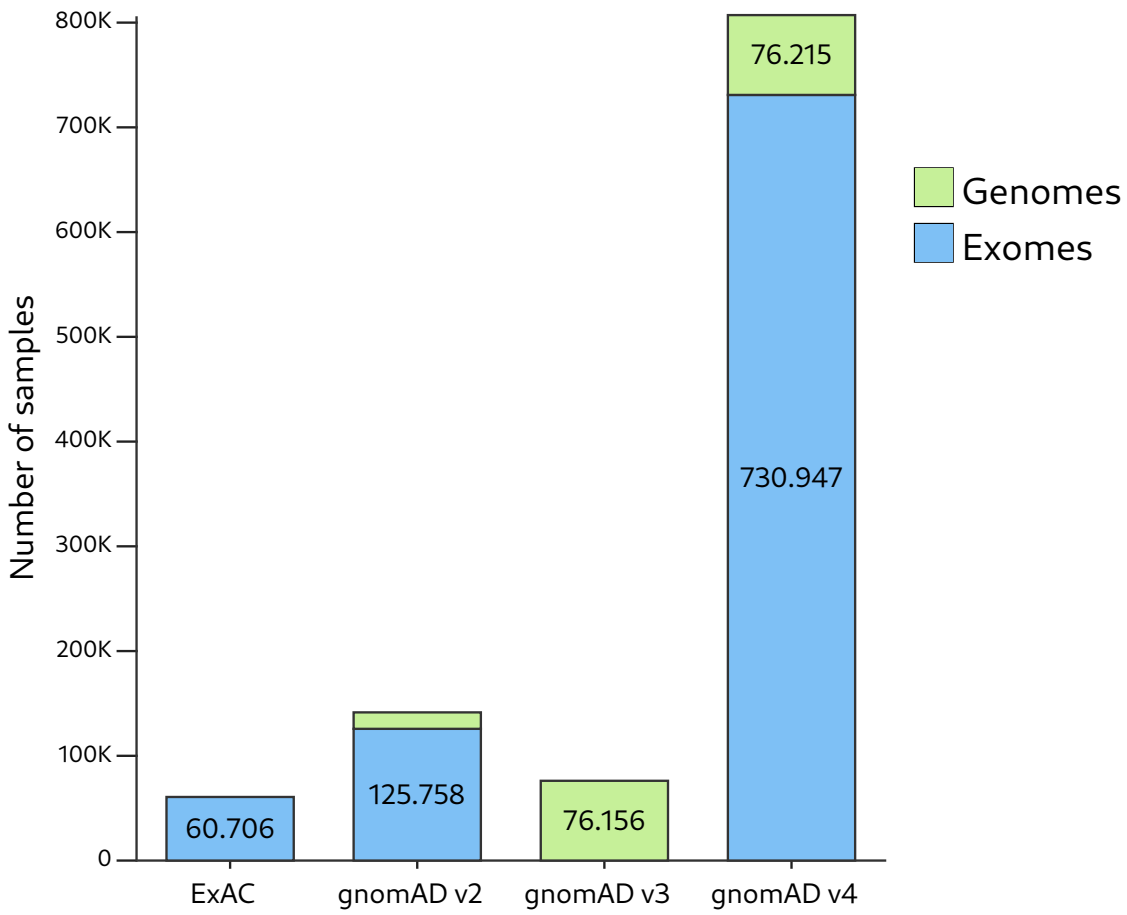

Figure 2.15: Sample size across major ExAC/gnomAD releases. Figure from [135] (public domain).

3 Benchmarking aberrant gene expression prediction in human tissues

3.1 Motivation

Aberrant gene underexpression is typically indicative of severely impaired or lost function. The development of algorithms prioritizing variants as potential genetic causes of gene expression outliers identified in RNA-seq samples has revealed that underexpression outliers are frequently associated with rare genetic variants (see section 1.3.3). However, before this work there was no algorithm predicting aberrant underexpression of genes based on an individual's genome. Also, the performance of existing variant annotation tools in the identification of gene underexpression outliers was not evaluated. To address this unmet need, I developed the first benchmark for predicting aberrantly expressed protein-coding genes in human tissues. Then I used this benchmark to evaluate the performance of existing variant annotation tools on predicting underexpression outliers (fig. 3.1). I focused specifically on underexpression outliers, as these can be expected to have severely impaired or lost function, whereas the functional consequence of an overexpression outlier is less clear and could potentially result in a gain of function.

3.2 A benchmark for tissue-specific aberrant expression prediction

To create the benchmark, I used gene expression outlier calls from the Genotype-Tissue Expression (GTEx) project. GTEx provides a large resource of whole-genome sequencing and RNA expression data from a wide range of organs and body parts (section 2.12.1). In its version 8 release, it provides more than 17,000 RNA-seq samples (see section section 2.4.2) of 948 assumed healthy individuals, collected postmortem from 54 different tissues. Most of these individuals also have paired whole-genome sequencing data (see section section 2.4) available.

I downloaded the GTEx RNA-seq read alignment files in the BAM format from dbGaP (phs000424.v8.p2). GTEx v8 provides variant calls for SNVs and small INDELs, but not for structural variants. Since previous studies showed that structural variants can potentially explain a large proportion of outliers (section 1.3.3), I selected 635 whole-genome sequences from GTEx version 7 (dbGaP entry phg000830.p1), an older release aligned to the GRCh37 reference genome for which structural variant calls are available

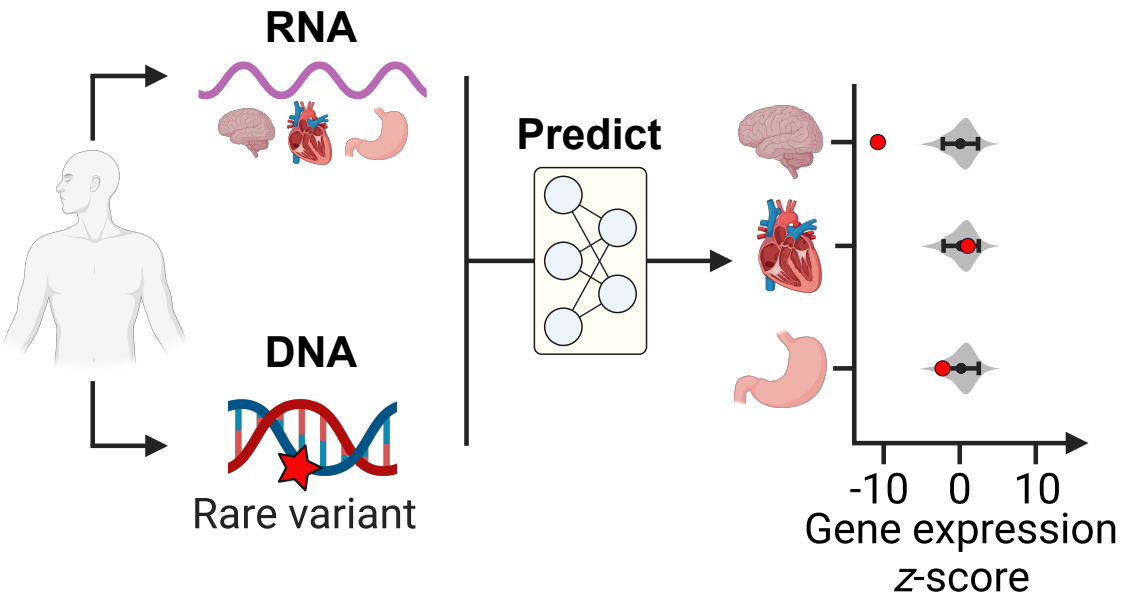

Figure 3.1: A benchmark for aberrant gene expression prediction across human tissues. The aim is to predict whether protein-coding genes are aberrantly underexpressed across human tissues based on DNA and, optionally, RNA-seq data of clinically accessible tissues. Therefore, I created a benchmark for aberrant underexpression prediction by processing 11,096 RNA-seq samples from 633 individuals across 48 tissues from GTEx. Figure created with BioRender.com.

from Ferraro and colleagues[39].

3.2.1 Expression outlier calling

As a first step, it was necessary to identify gene expression outliers in the GTEx dataset. To this end, Vicente A. Yépez applied the aberrant expression module of DROP v1.1.0[150] based on OUTRIDER (section 2.5)[14], separately for each tissue in the GTEx dataset. First, in every RNA-seq sample, the number of fragments (read pairs) per gene was counted by assigning a fragment to a gene if and only if the read pair was entirely aligned within the gene. This means that the same fragment might be assigned to more than one gene. Definitions for gene regions were taken from the GRCh38 primary assembly release 34 of the GENCODE project[45]. Genes with less than 1 fragment per kilobase of transcript per million mapped reads (FPKM, see section 2.4.2) in 95% or more of the samples in a tissue were considered insufficiently expressed in that tissue and were removed from further analysis. The aberrant expression module of DROP then applied OUTRIDER, a statistical method designed for identifying genes with abnormal expression in RNA-seq datasets. To ensure sufficient statistical power, we excluded tissues with less than 100 RNA-seq samples from OUTRIDER analysis. Therefore, six tissues were excluded from OUTRIDER analysis (bladder, endocervix, ectocervix, fallopian tube, kidney medulla, and kidney cortex). I further labeled all observations with a False Discovery Rate (FDR) less than 20% as gene expression outliers as this relaxed FDR cutoff of 20% turned out to help by leading to more robust evaluations and models. Using this FDR, OUTRIDER identified 42,727 underexpressed and 70,978 overexpressed genes as significant outliers, across 48 tissues from 946 individuals.

3.2.2 Filtering of expression outliers

To filter the expression outlier calls, I first subsetted them to protein-coding genes and removed all individuals without whole-genome sequencing data. Next, I tried to reduce the proportion of data points that could not be detected as outliers due to a lack of statistical power. An empirical investigation determined that, when applying an FDR cutoff of 5%, recovering half of the transcriptome-wide two-fold reduction outliers requires an expected fragment count ($\mu_{s,g}$) of at least 450[149]. Therefore, I removed all observations for which OUTRIDER reported an expected fragment count of less than 450. Also, I removed RNA-seq samples with more than 50 outliers. Samples with a high number of outliers could indicate cases where OUTRIDER was unable to fit the data accurately, or these are samples with globally affected gene expression, leading to widespread aberrant expression across the genome. Such expression aberrations are not predictable based on local sequence variation alone. Filtering samples with more than 50 outliers removed only 0.9% of samples (fig. 3.2), but reduced the total number of outliers by 10.6%.

The final dataset contained 17,637 underexpression outliers, 25,939 overexpression

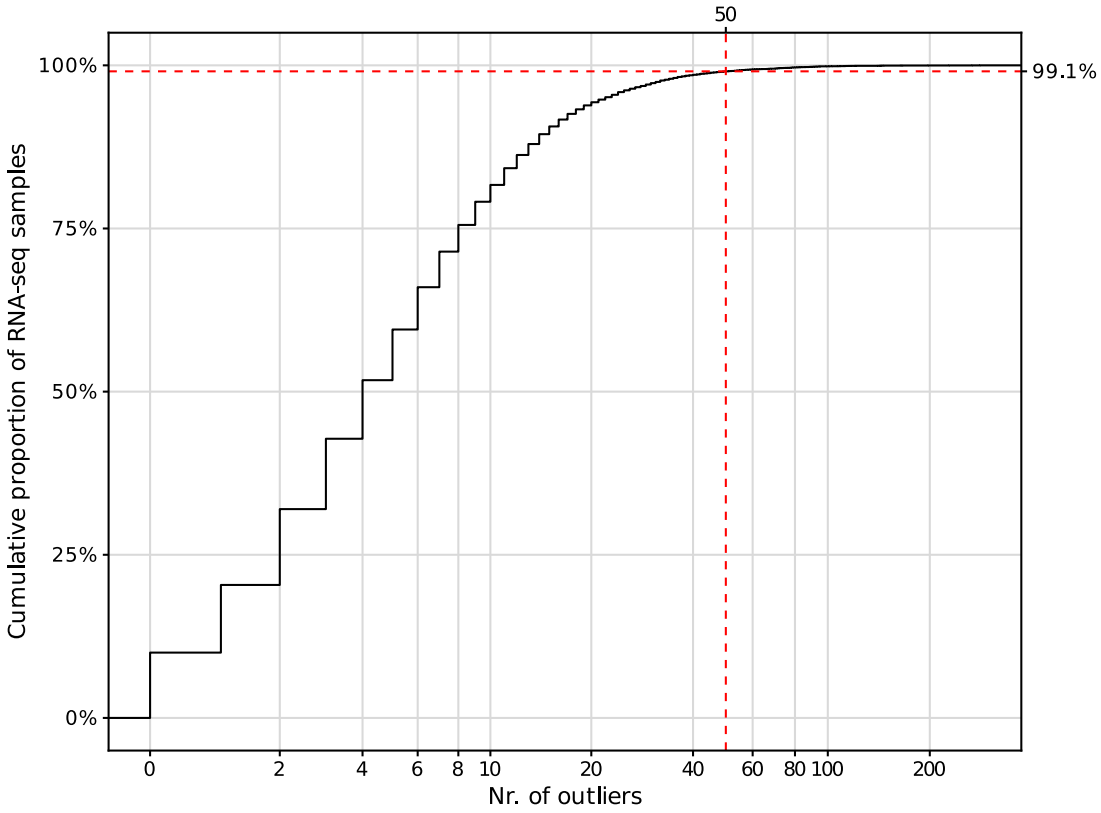

Figure 3.2: Only 0.9% of GTEx transcriptome sequencing samples have more than 50 outliers. The figure shows the cumulative proportion of RNA-seq samples (y-axis) that have at most a given number of outliers (x-axis) in the GTEx dataset. The vertical dashed line denotes the 50 outlier cutoff, which 99.1% of samples passed (horizontal dashed line).

outliers, and 100,281,961 non-outliers in 11,096 RNA-seq samples from 633 individuals across 48 tissues (table 3.1). This corresponds to about 1.6 underexpression and 2.3 overexpression outliers per RNA-seq sample. The class imbalance between outliers and non-outliers was large, with an outlier frequency of only 0.04% among observations. Underexpressed outliers appeared in 5,428 distinct genes and overexpressed outliers in 8,016 distinct genes. With a total of 1,744 samples, the majority of RNA-seq samples originate from brain tissues, closely followed by 1,219 samples from skin tissues (fig. 3.3). In contrast, the median number of samples per tissue type in the GTEx dataset is only 246.

90% of the overexpression outliers and 78% of the underexpression outliers are singletons, i.e. only aberrantly expressed in a single tissue (fig. 3.4). Further, most tissues have in median two overexpression and two underexpression outliers per sample, with median numbers ranging between one and three outliers of a certain kind (fig. 3.5a). This observation also holds when comparing samples across tissue types (fig. 3.5b).

The decision against the version 8 genome release of GTEx reduced the number of expression outliers by about 20%. However, this decision enabled the inclusion and consideration of structural variants in the benchmark.

3.2.3 Rare variant filtering

Given the low prevalence of expression outliers (0.04% of observations), I reasoned that variants with an allele frequency above 0.1% are unlikely to cause an expression outlier. Therefore, I removed SNPs and short INDELs if they had a minor allele frequency in the general population ≥ 0.001 based on the Genome Aggregation Database (gnomAD v2.1.1, see section 2.12.6) and were found in at least 2 individuals within GTEx. I also removed variants that were supported by less than 10 reads and did not pass a conservative genotype-quality filter of GQ > 30. For structural variants, I only filtered for the number of occurrences in the GTEx dataset and kept those present in less than two individuals.

Further, I focussed on cis-regulatory variants by considering variants located within the gene and up to 5,000 bp around the gene to fully cover promoter and transcription termination regions, resulting in a total of 8.2 million rare variants. 59% (10,429) of underexpression outliers and 43% (11,144) of overexpression outliers harbor rare variants within $\pm$ 5,000 bp of the gene (fig. 3.6a). Therefore, with this set of variants, the maximum achievable recall is limited to 59% of the underexpression outliers. Considering only singleton outliers, i.e. genes in individuals that are aberrantly expressed in a single tissue only, 40% of the overexpression singletons and 44% of the underexpression singletons harbor rare variants with 5,000 bp of the gene (fig. 3.6b).

Filtering step	Individuals	Genes	Tissues	Samples	Under-expressed outliers	Non-outliers	Over-expressed outliers
Unfiltered	946	33520	48	16146	42727	283989323	70978
Samples with whole genomes	633	33520	48	11174	29915	196304177	49864
Keep only protein-coding genes	633	18549	48	11174	23858	147294919	41410
Remove samples with many outliers	633	18549	48	11096	22669	146254583	36335
Keep only genes of samples that have sufficiently large expected number of reads ($\mu > 450$)	633	18152	48	11096	17637	100281961	25939

Table 3.1: GTEx dataset filtering.

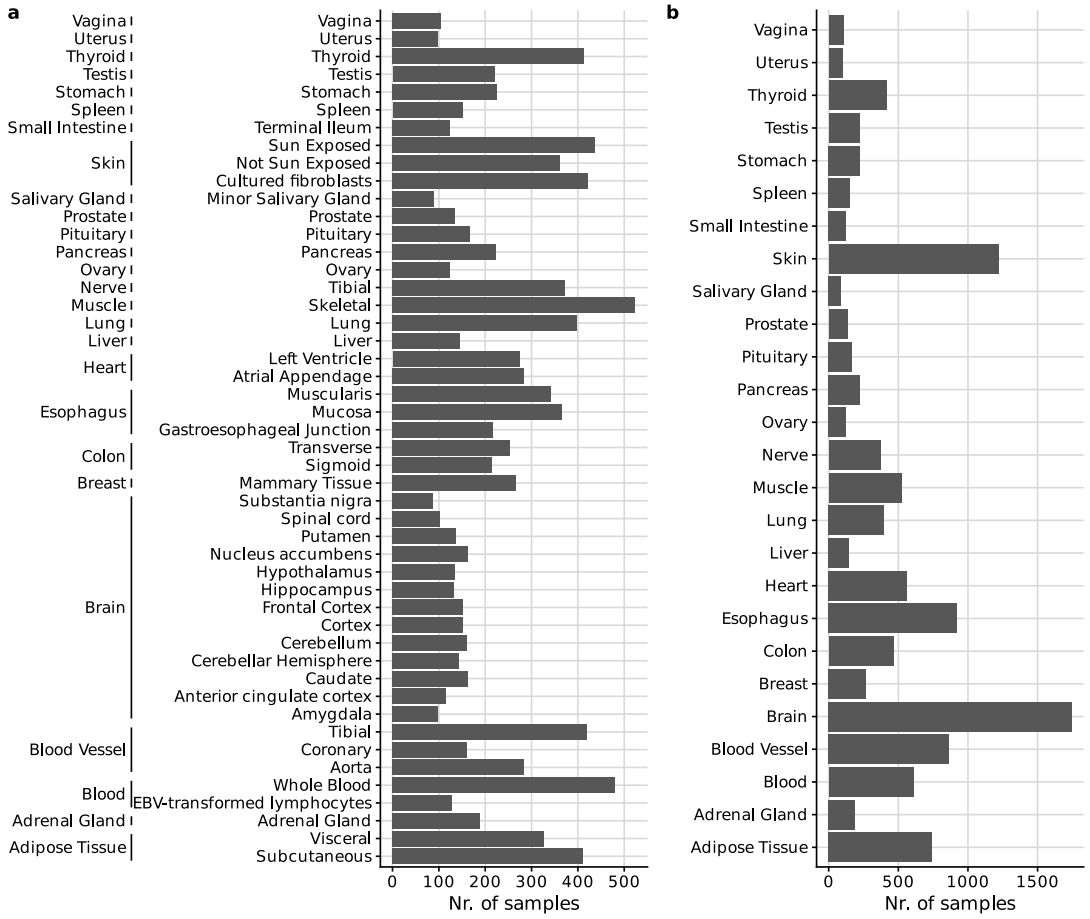

Figure 3.3: Most RNA-seq samples in the GTEx dataset are from skin or brain tissues. The figure shows the number of samples per tissue **(a)** and tissue type **(b)**. In total, there are 11,096 RNA-seq samples in the GTEx dataset.

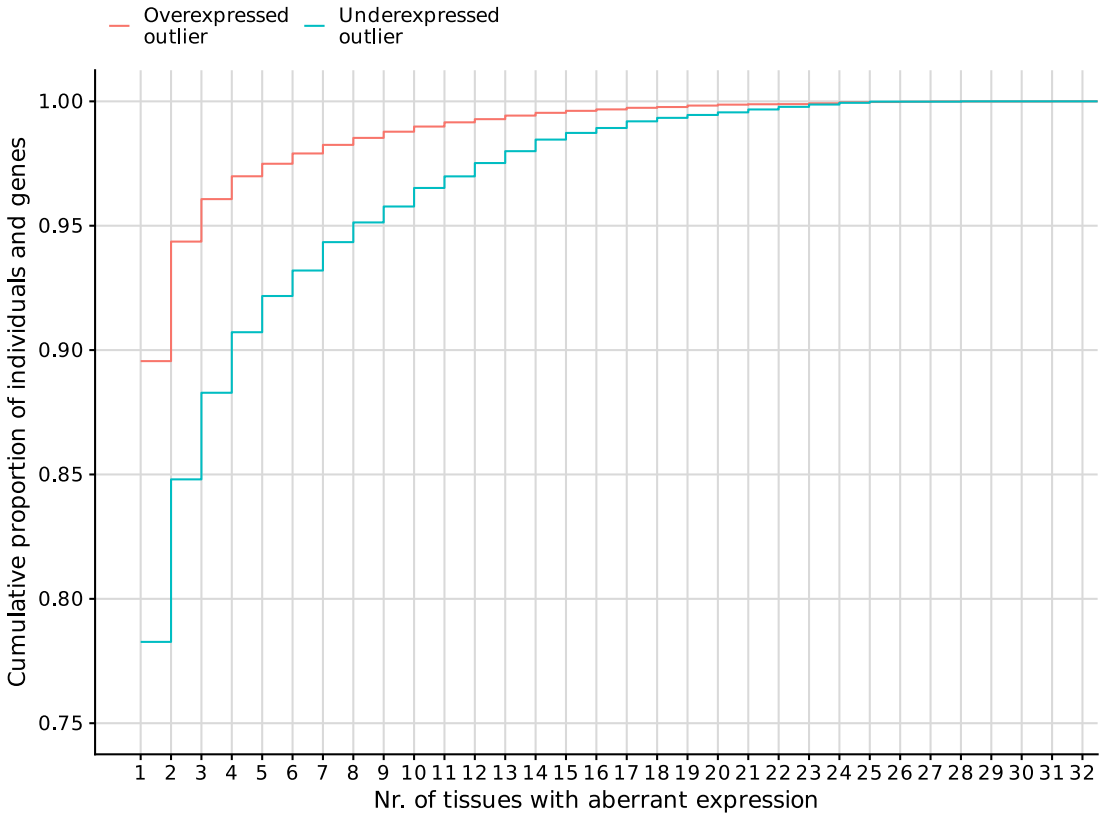

Figure 3.4: 90% of the overexpression outliers and 78% of the underexpression outliers are singletons. The figure shows the cumulative proportion of individuals and genes (y-axis) that are an outlier in at most a given number of tissues (x-axis) in the GTEx dataset.

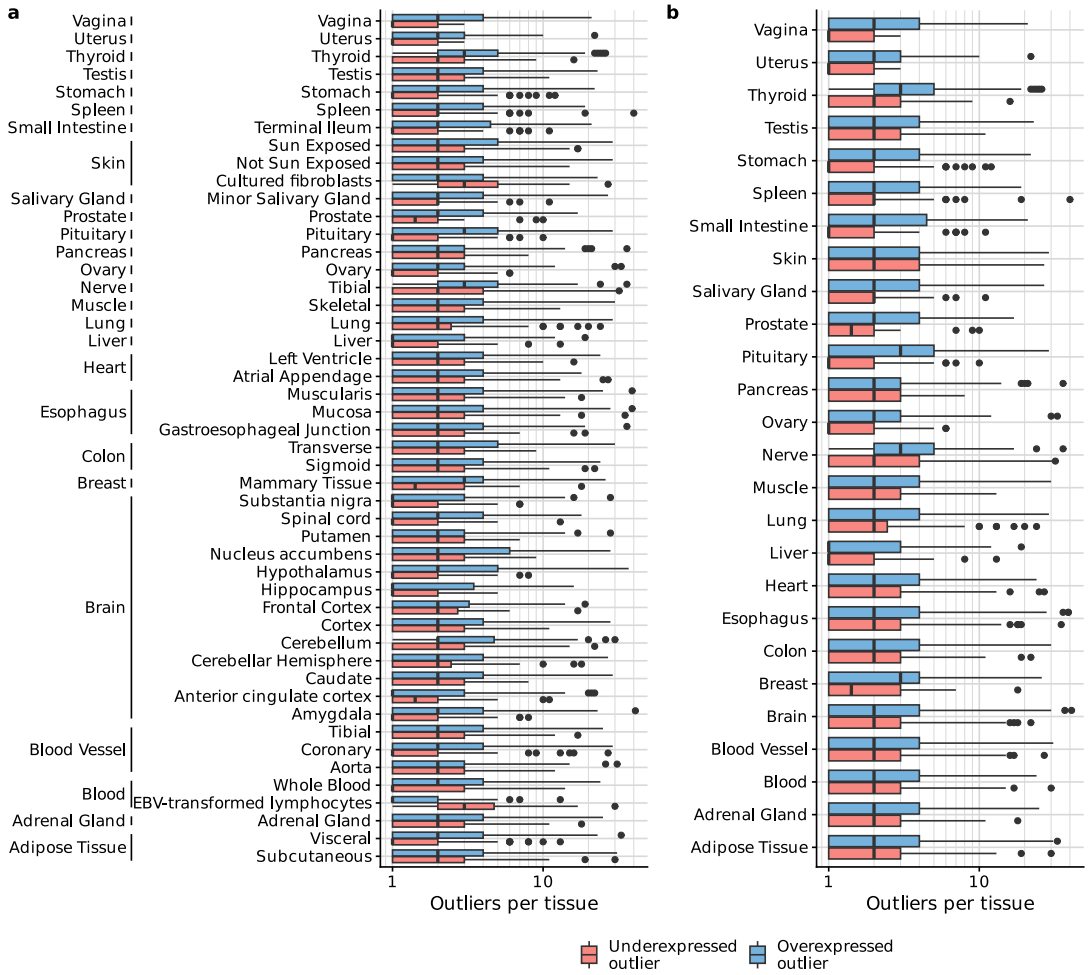

Figure 3.5: In most tissues, samples have in median two underexpression and two overexpression outliers. The figure shows the outliers per tissue **(a)** and tissue type **(b)** in the GTEx dataset, colored by the outlier type.

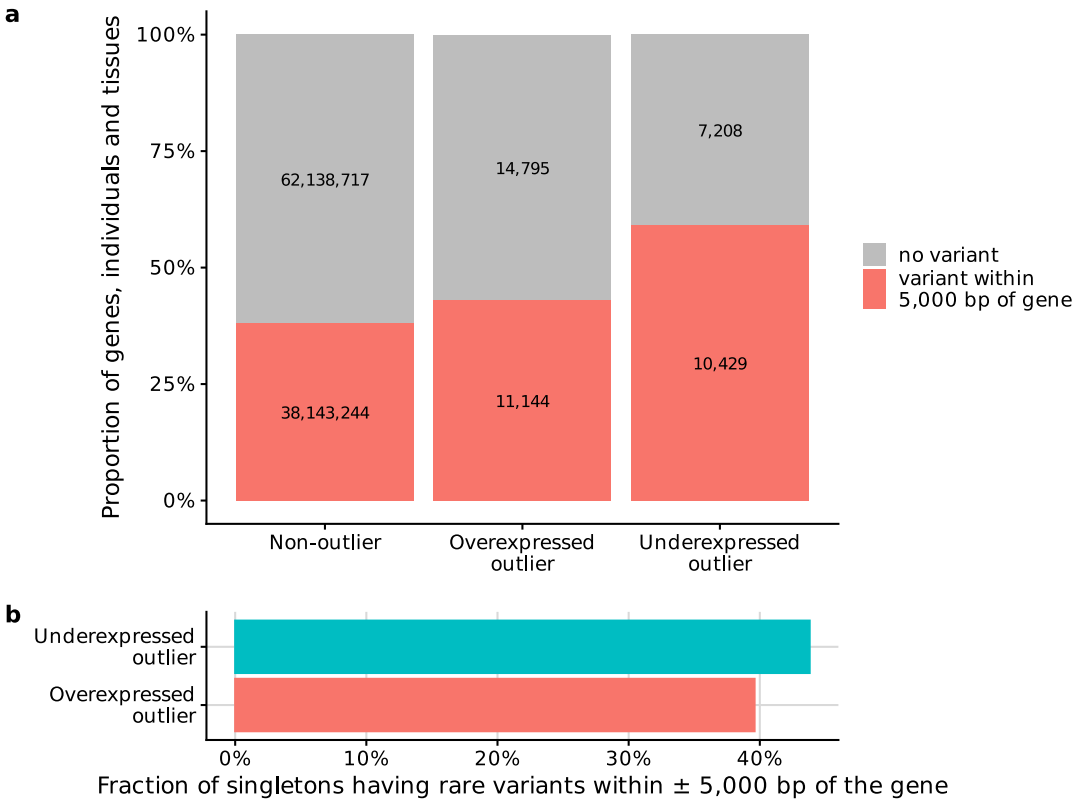

Figure 3.6: Not all outliers can be explained by variants within gene regions. **(a)** Proportion of overexpression outliers, underexpression outliers and non-outliers harboring rare variants within ± 5,000 bp of the gene. **(b)** Fraction of singletons harboring rare variants within ± 5,000 bp of the gene.

3.2.4 Benchmark task and evaluation metric

The benchmark task is to predict whether some protein-coding gene is aberrantly underexpressed in a certain human tissue, based on the rare variants of the individual. Due to the large class imbalance in the expression outlier benchmark dataset, I chose to evaluate models using precision-recall curves and summarize them with the area under the precision-recall curve (AUPRC, see section 2.6).

3.3 Performance of variant annotation tools in aberrant expression prediction

After creating the benchmark dataset, I set out to test whether existing variant annotation tools that were not specifically designed to predict aberrant underexpression could provide informative signals.

3.3.1 Enrichment of variant consequences

As a first step, I evaluated whether certain types of variant consequences are enriched among expression outliers. To this end, I first calculated the consequences of all rare variants using Ensembl VEP (section 2.9) v108 and removed any annotations that do not affect the canonical transcript of a gene, as annotated by VEP. Then, I determined the proportion of genes that are affected by variants with a certain consequence among outlier classes (fig. 3.7a).

Among underexpression outliers, I discovered a strong enrichment of frameshifts, variants affecting start and stop codons, and splicing variants which are associated with triggering nonsense-mediated decay of the transcript (section 1.3). Another category of variants frequently observed among underexpression outliers were variants affecting the 5' UTR, a region that is important for transcription initiation (section 1.3).

The enrichment of missense variants and 3' UTR variants was less strong. Missense variants affect the protein sequence without changing its length. Although missense variants mainly affect the functionality of the protein, changes in the encoded protein sequence can also lead to ribosomal stalling and mRNA degradation by the ribosome-mediated quality control and therefore a reduction in gene expression levels[125]. The 3' UTR contains signals that are important for transcription termination (section 1.3). However, while missense variants and 3' UTR variants were found more frequently among underexpression outliers than in other classes, these types of variants were also present in many overexpression outliers and non-outliers.

A particularly strong enrichment for underexpression outliers was found in transcript ablations, which were inferred from the structural variant deletion calls. Out of 320 observations harboring transcript ablations from 43 GTEx individuals, 199 observations from 33 individuals were underexpression outliers. None of the overexpression outliers were harboring a transcript ablation. Transcript ablations typically delete the whole

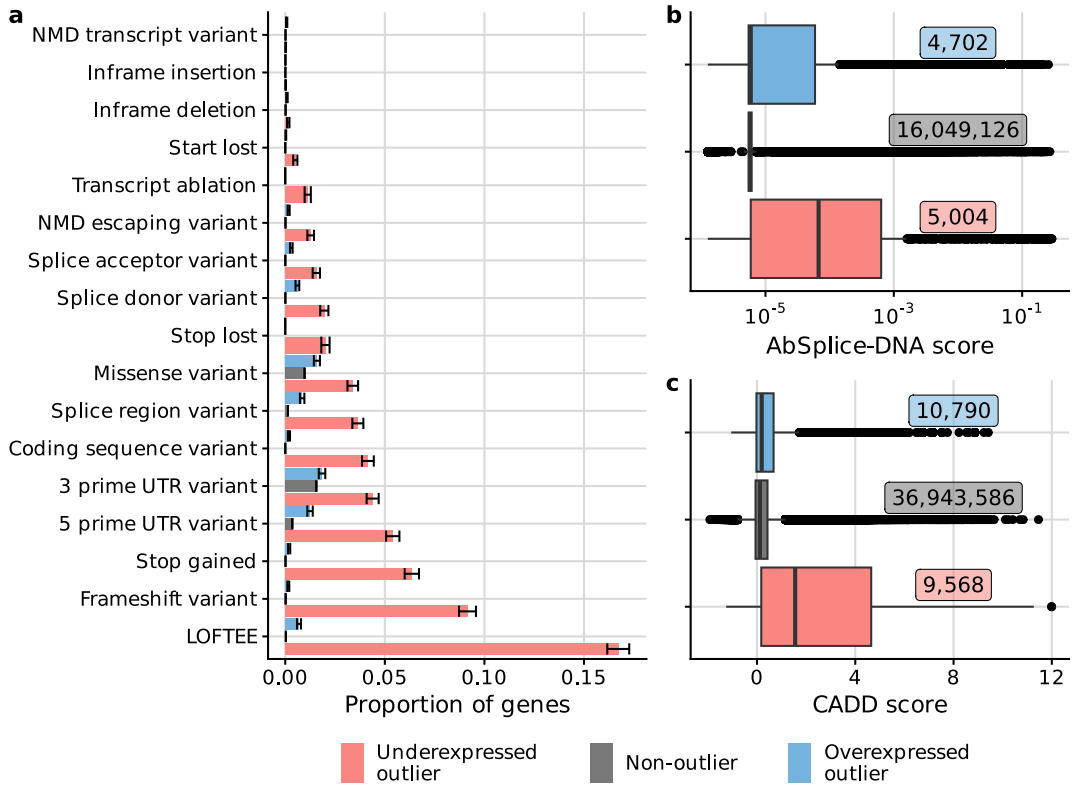

Figure 3.7: Enrichment of various variant annotations in expression outliers across tissues. In total, there are 17,637 underexpressed genes (red), 25,939 overexpressed genes (blue), and 100,281,961 non-outliers (gray) across all tissues. **(a)** Proportion of underexpression outliers, overexpression outliers, and non-outliers with a rare variant of a given annotation. Error bars mark 95% binomial confidence intervals. **(b)** Distribution of gene-level CADD scores among genes with CADD-annotated variants. **(c)** Distribution of gene-level AbSplice scores among genes with AbSplice-annotated variants. **For all boxplots:** Box label, sample size; center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

transcript from the sequence and therefore are expected to have a large impact on gene expression.

Overall, these findings align with previous studies (see section 1.3.3).

3.3.2 LOFTEE

Furthermore, I explored the use of LOFTEE, a tool designed to predict a high-confidence subset of loss-of-function variants, particularly those likely to induce nonsense-mediated decay. LOFTEE implements various filters, such as excluding stop-gained and frameshift variants that are within 50 bp of the end of the transcript as these typically escape nonsense-mediated decay, or variants that only affect splicing in untranslated regions.

To generate LOFTEE annotations for rare variants, I used the LOFTEE plugin for VEP and removed all LOFTEE variants affecting non-canonical transcripts (canonical as annotated by VEP).

In the GTEx dataset, I found that over 17% of aberrantly underexpressed genes had a LOFTEE-positive, a stark contrast to non-outliers where less than 1% had a LOFTEE variant (fig. 3.7a).

3.3.3 AbSplice

The splice sites employed by LOFTEE and VEP are derived exclusively from the genome annotation. Consequently, these tools do not take into account any splice sites that are not annotated in the reference genome. However, standard genome annotations are not tissue-specific and the splicing events can vary significantly between different tissues or developmental stages[9]. Also, many weak splice sites, i.e. sites that are spliced at a low level, are missing from standard genome annotations. However, these sites can be activated by genetic variants, leading to the formation of novel exons[26, 82].

AbSplice is a recent tool that predicts aberrant splicing across tissues using a more comprehensive map of splice sites, including unannotated weak splice sites, and their tissue-specific usage[144]. I contributed to the development of AbSplice in data curation, validation, formal analysis, and visualizations. For a comprehensive and logically articulated thesis, I here summarize the results of AbSplice.

Using 16,213 RNA-seq samples of the Genotype-Tissue Expression (GTEx) dataset, spanning 49 tissues and 946 individuals, we established a comprehensive benchmark for predicting variants leading to aberrant splicing in human tissues, spanning over 8.8 million rare variants (fig. 3.8a). We then evaluated the predictive performance of two state-of-the-art sequence-based splicing models, MMSplice[20] and SpliceAI[66](fig. 3.8b). MMSplice predicts quantitative usage changes of predefined splice sites within a 100-bp window of a variant. SpliceAI predicts the creation or loss of splice sites within a 50-bp window of a variant, independent of gene annotations. The performance evaluation revealed limited performance of both models, with an overall precision of 8% for MMSplice and of 12% for SpliceAI at 20% recall (fig. 3.9).

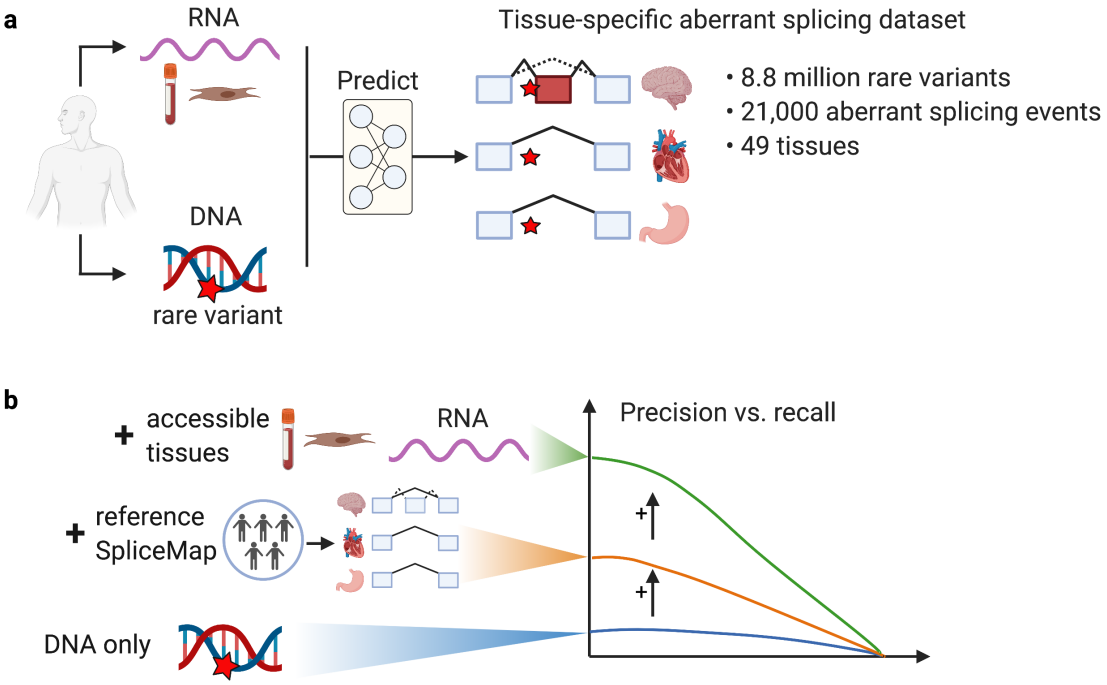

Figure 3.8: Study design and main findings of AbSplice. The aim is to predict whether rare variants are associated with aberrant splicing across 49 human tissues. **(a)** A comprehensive benchmark for aberrant splicing was established by processing GTEx samples with a recently published aberrant splicing caller[105]. Based on this benchmark, predictors could be assessed and developed that take as input DNA sequence and, optionally, RNA-seq data of clinically accessible tissues. **(b)** Benchmarking revealed modest performance of currently used algorithms based on DNA only, a substantial performance improvement when integrating these models with SpliceMap, a quantitative map of tissue-specific splicing we developed in this study, and further improvements when also including direct measures of aberrant splicing in accessible tissues. Figure created with BioRender.com.

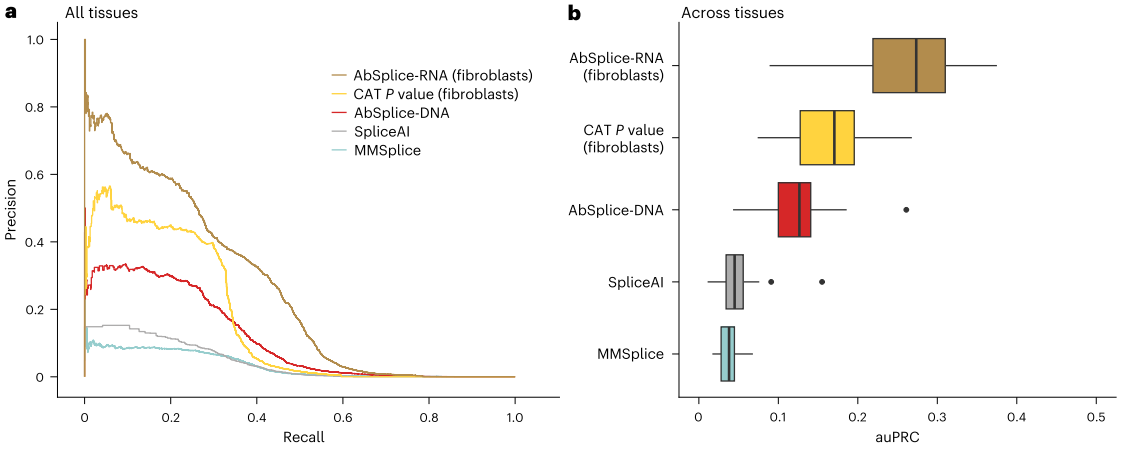

Figure 3.9: AbSplice outperforms state-of-the-art splicing models in predicting aberrant splicing. **(a)** Precision-recall curve comparing the overall prediction performance on all GTEx tissues of SpliceAI, MMSplice using GENCODE annotation, AbSplice-DNA, gene-level aberrant splicing *P*-values in fibroblasts, and AbSplice-RNA, which integrates AbSplice-DNA features with features from RNA-seq from fibroblasts. **(b)** Distribution of the AUPRC of the models in c across tissues (n = 49). Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles; P values were computed using the paired one-sided Wilcoxon test.

Next, we set out to improve on predicting aberrant splicing across human tissues (fig. 3.8b). Based on GTEx, we created a tissue-specific splicing annotation (SpliceMap) of acceptor and donor splice sites and their usage (Ψ_{ref}, reference level of the percent spliced-in value for a particular splicing site) in 49 human tissues. We then trained AbSplice-DNA, a model integrating SpliceMaps with MMSplice and SpliceAI, to predict aberrant splicing across tissues. This led to a threefold increase of precision at the same recall, with a significantly higher AUPRC consistently across tissues (fig. 3.9). These performance increases replicated in two independent cohorts.

Additionally, we found that RNA-seq from clinically accessible tissues complements DNA-based splicing predictions when incorporated into an integrative model AbSplice-RNA, increasing precision to 60% at 20% recall.

While investigating false positive predictions, we suspected that some of these might actually be correct. When aberrant splicing isoforms trigger nonsense-mediated mRNA decay, these isoforms barely have any reads in RNA-seq data and hence are typically not detected by aberrant splicing callers. However, the extent to which aberrant splicing causes nonsense-mediated mRNA decay was not clear at that time. Therefore, I now investigated the enrichment of gene-level AbSplice-DNA scores, generated by Nils Wagner, among underexpression outliers in GTEx.

I found that AbSplice-DNA scores were in median 10 times higher among underexpression outliers than among non-outliers and overexpression outliers (fig. 3.7b), suggesting an effect of aberrant splicing on aberrant gene expression. About 28% (5,004) of the

underexpression outliers harbored an AbSplice-DNA annotated rare variant.

3.3.4 CADD

Another tool that could potentially be predictive of expression outliers is CADD[77, 120, 121], a tool to estimate the deleteriousness of a genetic variant. It was trained on a binary distinction between simulated de novo variants and variants that have arisen and become fixed in human populations (see section 2.11). The advantage of CADD is that it can score any possible human SNVs or small INDELs.

To obtain CADD scores, I applied the CADD plugin of VEP to all SNVs and INDELs in the GTEx dataset and max-aggregated all variant scores within ± 5,000 bp of each gene. While 10,429 underexpression outliers had at least one rare variant within this range, the CADD plugin only reported scores for 9,568 genes. This deviation is due to some variants not being available in the precomputed set of CADD scores that the VEP plugin uses to annotate variants.

I found that the gene-level CADD scores are in the median about 13 times higher in underexpression outliers than in non-outliers and overexpression outliers (fig. 3.7c).

3.3.5 Performance comparison of LOFTEE, AbSplice, and CADD on aberrant underexpression prediction

In GTEx, all three tools show mild predictive performance (fig. 3.10). LOFTEE-positive variants recalled 16.7% of the underexpression outliers at a precision of 9.7%, with a median AUPRC across tissue types of 1.1%. CADD, with a median AUPRC of 0.9%, never passes this level of precision. At a low recall rate of 0.1%, AbSplice reaches a peak precision of 31.6%, suggesting that the most extreme splicing outliers indeed can lead to aberrant underexpression of genes. However, AbSplice is overall the worst-performing model, reaching only 0.5% AUPRC in median across tissue types.

3.4 Summary

In conclusion, I created a new benchmark for aberrant expression prediction based on GTEx. I found that the variant annotation tools LOFTEE, AbSplice and CADD show strong enrichment in underexpression outliers. However, the predictive value of these models is limited.

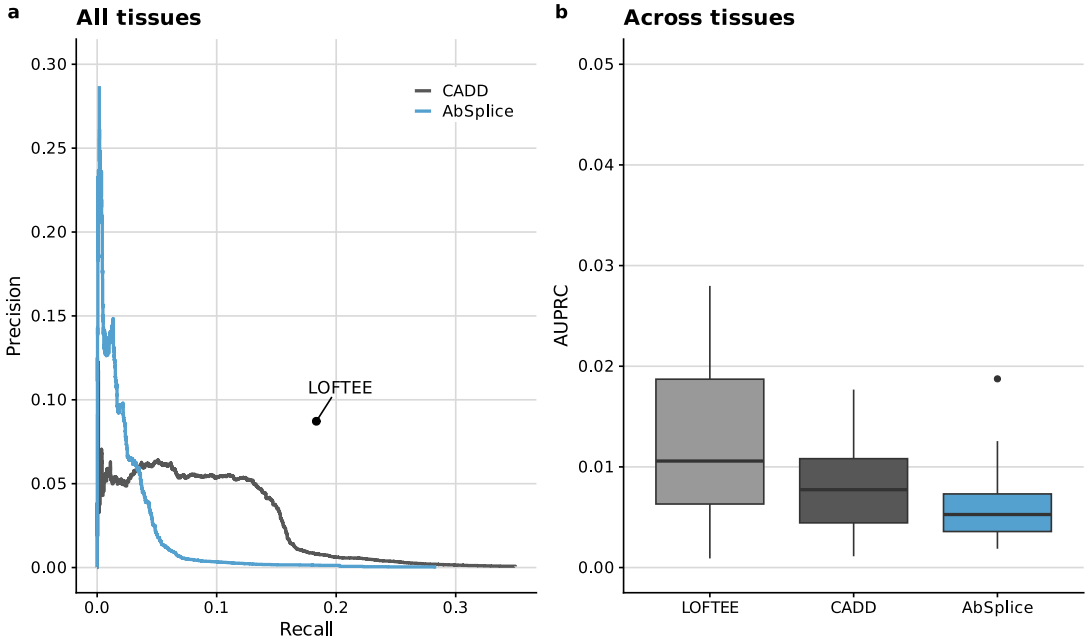

Figure 3.10: Performance of various variant annotations in underexpression prediction. **(a)** Precision-recall curve for all tissues combined. LOFTEE shows up as a single point because it is a binary filter. **(b)** Distribution of average precision (AUPRC) across 26 GTEx tissue types. P-values were obtained using a paired Wilcoxon test. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

4 AbExp: Predicting aberrant gene underexpression across human tissues

4.1 Motivation

Given the limited performance of existing variant annotation tools, I wanted to develop a specialized method that can improve the prediction of underexpression outliers. In this section, I will describe the development of AbExp, a model combining various variant and tissue annotations to significantly improve the prediction of aberrant underexpression outliers across 48 human tissues based on genetic variants (fig. 4.1).

4.2 Training and evaluation procedure

One observation in the GTEx dataset is a combination of a gene in an individual and tissue. The aim is to predict whether this observation is an underexpression outlier, based on a set of (gene-level) features. To avoid overfitting on individual-specific variants, I split all 633 individuals of the GTEx dataset into six cross-validation groups with approximately equal numbers of underexpression outliers and tissues (fig. 4.2). To train and evaluate a prediction model, the model was trained six times on five of these folds and evaluated on the held-out fold, each time using a different held-out fold as the validation set. Precision recall curves were then derived from the predicted scores of the six validation folds. AUPRC distributions were calculated across 26 tissue types to group together highly similar tissues, notably many regions of the brain (see also fig. 3.3). This grouping by tissue type avoids reporting inflated performance driven by a set of highly similar tissues.

The here described training and evaluation procedure was used for all the models presented in this chapter.

4.3 Integrating rare variant annotations to predict underexpression outliers across tissues

I started with training a simple tissue-independent model integrating CADD, LOFTEE, and VEP consequences of rare variants. I did not include transcript ablations inferred from structural variants as structural variant calls are often not available and I wanted to investigate their effect separately.

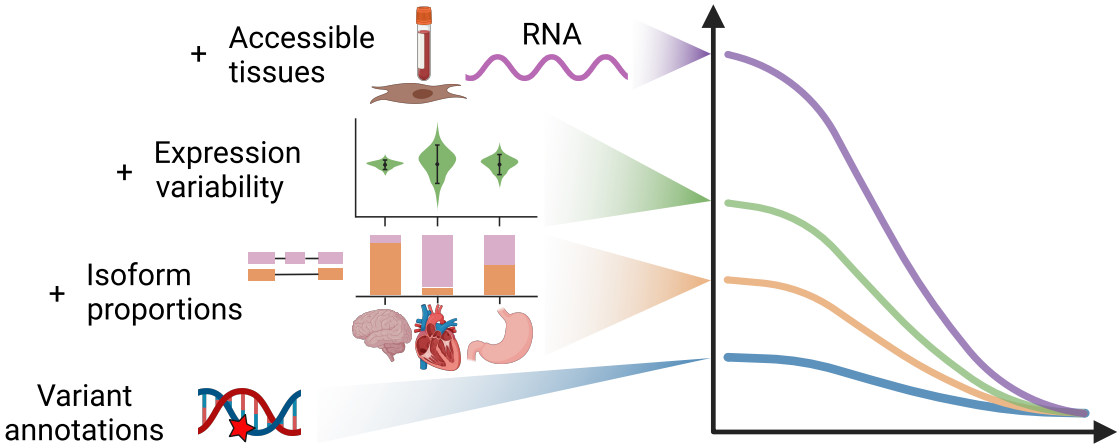

Figure 4.1: Improving aberrant underexpression prediction in human tissues. By assessing various variant and tissue annotations, it became evident that predictions could be significantly enhanced by weighting variant effects with tissue-specific isoform proportions and incorporating the expression variability of a gene. Further integration of expression measurements from clinically accessible tissues led to another two-fold improvement. Figure created with BioRender.com.

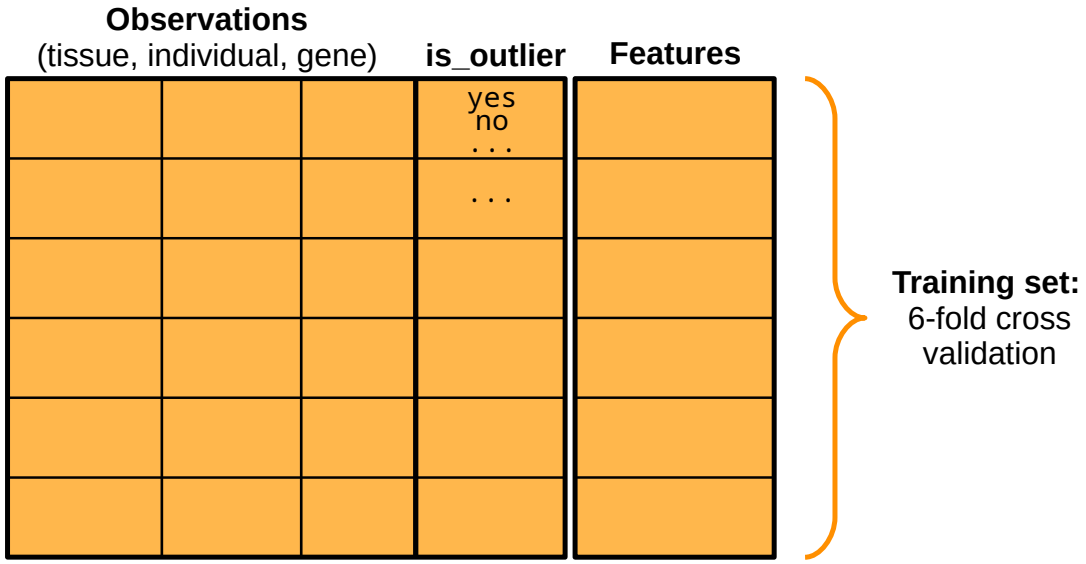

Figure 4.2: Cross-validation scheme. One observation in the dataset is a combination of a gene in an individual and tissue. The aim is to predict whether this observation is an underexpression outlier, based on a set of (gene-level) features. The whole dataset is split into six cross-validation folds with approximately equal numbers of underexpression outliers and tissues.

4.3.1 Calculation of gene-level features

Following the same procedure as in chapter 3, I used Ensembl VEP with LOFTEE and CADD plugins to annotate consequences, identify LOFTEE-positive variants, and predict CADD scores of all rare variants within 5,000 bp of genes. Then I removed any annotations that do not affect the canonical transcript of a gene (canonical as annotated by VEP).

4.3.2 Quantitative prediction of outlier state

Using these features, I trained a non-linear regression model to quantitatively predict the gene expression z-score, a value describing how many standard deviations the observed expression level deviates from the average population. While the prediction task is a binary classification of underexpression outliers, predicting the underlying standard-normal distributed z-scores (fig. 4.3) led to better models, as it allowed to overcome the large class imbalance between underexpression outliers and non-outliers by learning moderate effects that do not necessarily lead to significant outliers.

To obtain the gene expression z-scores, I quantile-mapped the OUTRIDER-fitted fragment count distributions (eq. (2.4)) to the standard normal distribution as follows:

$$z\text{-score} = \mathrm{CDF}^{-1}_{N(0,1)}\left(\mathrm{CDF}_{NB}(x|\mu,\theta)\right) \tag{4.1}$$

, where $\mathrm{CDF}^{-1}_{N(0,1)}$ is the inverse cumulative distribution function of the standard normal distribution and $\mathrm{CDF}_{NB}(x|\mu,\theta)$ the negative-binomial cumulative distribution function. An illustration of this quantile mapping can be seen in fig. 4.4.

As model architecture, I used a non-linear gradient-boosted decision trees model [59] from the LightGBM[74] framework with default parameters (see section 2.7.4).

Evaluating the performance of this model on held-out data showed that the predicted z-scores consistently outperformed ranking based on CADD scores (fig. 4.5a). Furthermore, the integrative model achieved the same precision at the same recall as filtering by LOFTEE variants, with the added benefit of providing a continuous score that allows the selection of more stringent cutoffs to yield a higher precision (up to 13%, fig. 4.5a). Across tissue types, this first integrative model significantly outperformed CADD and LOFTEE according to the average precision (fig. 4.5b).

4.4 Accounting for tissue-specific isoform expression

As described, the predictions of this first integrative model were tissue-independent: LOFTEE, CADD and VEP consequences provide the same effect annotation for a given transcript isoform, regardless of the tissue. Also, the considered canonical transcript isoforms were not tissue-specific. Therefore, the feature set for a gene of an individual does not differ between tissues and, consequently, neither do the predictions of this first

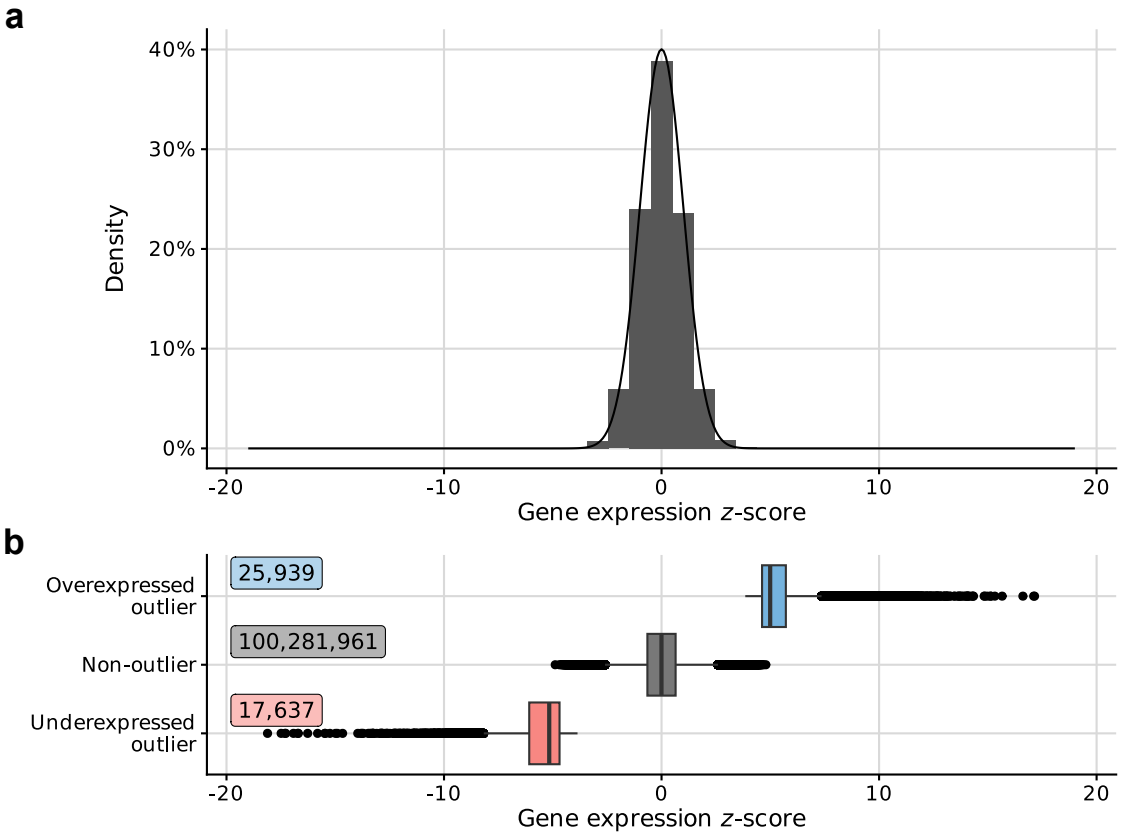

Figure 4.3: Gene expression *z*-scores are standard-normally distributed. **(a)** Histogram of 100,325,537 measured *z*-scores. The black curve shows the density of a standard-normal distribution. **(b)** Boxplot of measured *z*-scores among overexpressed outliers, non-outliers, and underexpressed outliers.

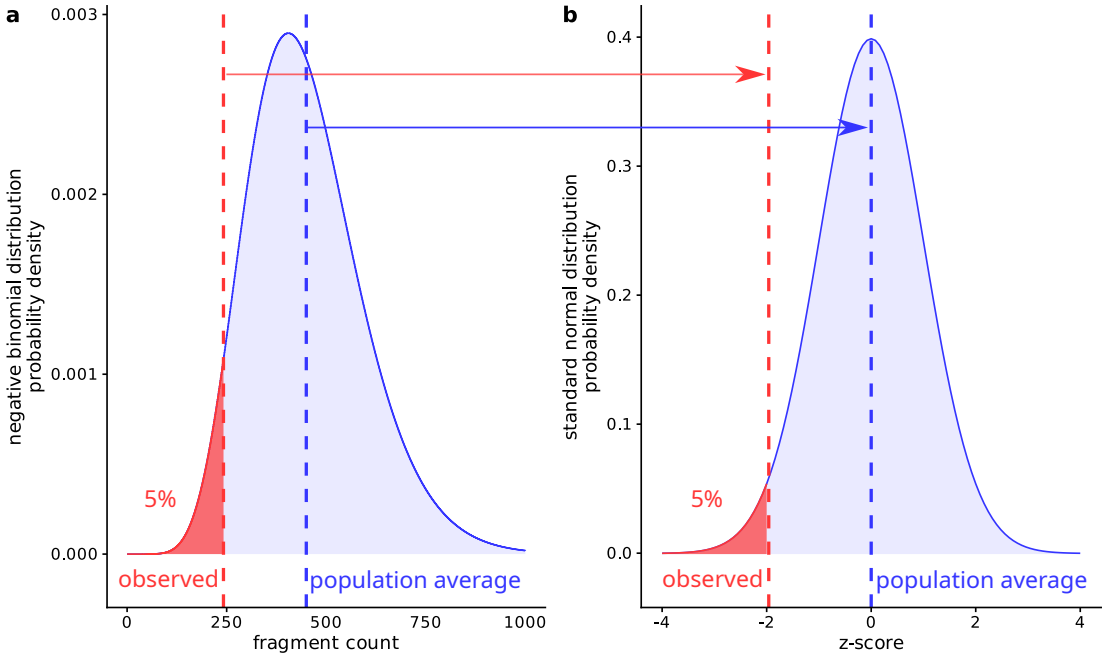

Figure 4.4: Quantile-mapping of OUTRIDER-fitted fragment count distribution to standard normal distribution. (a) Example of OUTRIDER-fitted fragment count distribution with a population average (blue dashed line) of $\mu = 450$ and a dispersion of $\theta = 10$. An observed fragment count of 242 (red dashed line) or less would be estimated to be present in less than 5% of the population (red area). **(b)** Observed (red dashed line) and expected (blue dashed line) fragment counts from (a) mapped to the standard normal distribution. The red area marks 5% of the population having an observed fragment count of 242 or less. This translates to a z-score of -1.96.

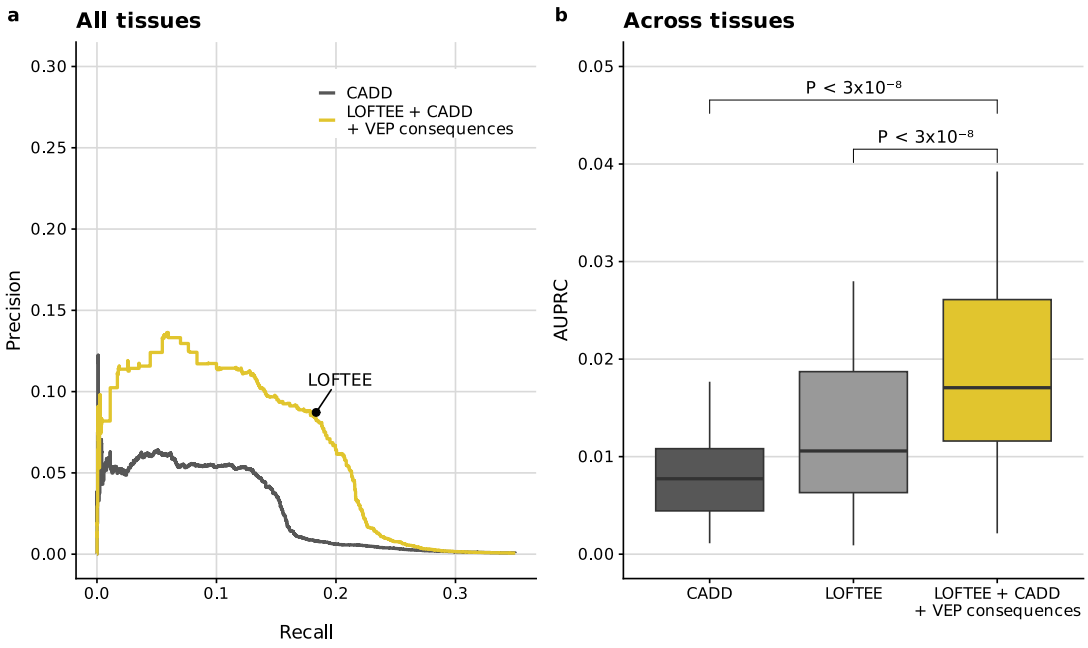

Figure 4.5: An integrative model outperforms both LOFTEE and CADD on predicting underexpression outliers. (a) Precision-recall curve for all tissues combined. LOFTEE shows up as a single point because it is a binary filter. **(b)** Distribution of average precision (AUPRC) across 26 GTEx tissue types. P-values were obtained using a paired Wilcoxon test. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

integrative model. However, variants can have tissue-dependent effects as the transcript isoforms of a gene are often expressed at different proportions across tissues[25].

4.4.1 Calculation of transcript isoform proportions in each tissue

To investigate the effect of tissue-specific isoform expression in GTEx, I first estimated the expression proportion of an isoform i in a tissue t as the median TPM proportion across individuals among all isoform of the same gene g:

$$\text{proportion}(i,t) = \text{median}_{s \in \text{individuals}} \left(\frac{\text{tpm}(i,t,s)}{\sum_{x \in \text{isoforms}(g)} \text{tpm}(x,t,s)} \right) \tag{4.2}$$

, with $\text{tpm}(i,t,s)$ as the transcript-level TPM obtained from GTEx v8 (dbGaP Accession phs000424.v8.p2).

An example for the impact of tissue-specific isoform expression in GTEx can be seen in fig. 4.6. Here, ENST00000358514, the canonical transcript of *PSMB10* according to the MANE annotation[107], was estimated to generate only about 4% of *PSMB10* total gene expression in putamen. The vast majority (91%) of *PSMB10* gene expression in putamen was attributed to another transcript, ENST00000570985. Conversely, in fibroblasts, the canonical transcript contributed to nearly 48% of the total gene expression. Exon 4 is not included in the transcript ENST00000570985 but is included in the canonical transcript ENST00000358514, explaining why a frameshift variant in exon 4 was associated with a high impact on gene expression in cultured fibroblasts but showed a limited effect in putamen (fig. 4.6a,b).

When investigating the amount of total gene expression contributed by canonical transcripts (using MANE-select[107] as the canonical transcript definition) across genes and tissues, I found that, in general, only 30% of the canonical transcripts contributed to more than 90% of the total expression of their gene and as much as 18% of the canonical transcripts contributed to less than 10% of their gene's total expression (fig. 4.7). Therefore, when considering only the variant consequence assigned to a single transcript isoform, relevant information is lost even if the isoform is annotated as the canonical one.

4.4.2 Calculation of gene-level features

To address this issue, I created a tissue-specific feature set by weighting isoform-specific variant annotations such as the VEP consequences and LOFTEE classification by the proportion of affected isoforms i per gene g and tissue t:

$$w_v(t) = \sum_{i \in \text{isoforms}(g)} \text{proportion}(i,t) \cdot \delta_{v \text{ affects } i} \tag{4.3}$$

, where $\delta_{v \text{ affects } i}$ is 1 if the variant affects the isoform, otherwise 0, and $\text{proportion}(i,t)$ as defined in eq. (4.2). All resulting variant annotations were then max-aggregated per

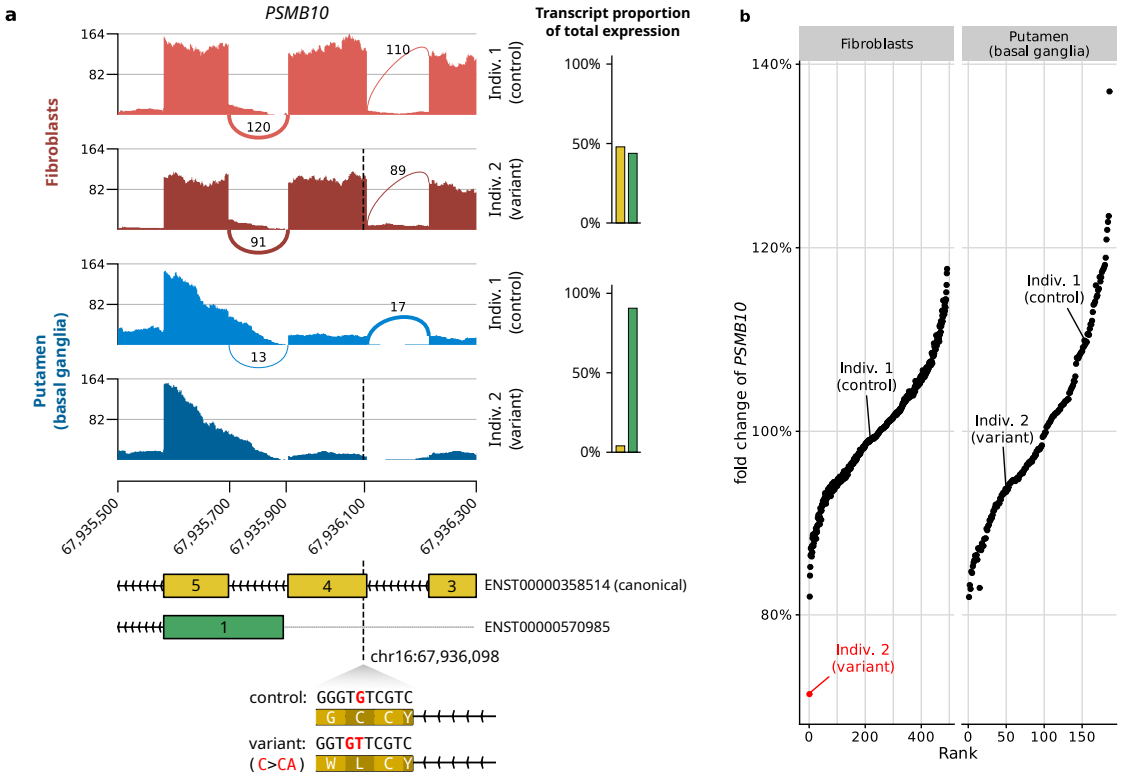

Figure 4.6: Tissue-specific isoform expression in *PSMB10* leads to tissue-specific aberrant gene expression. **(a)** Sashimi plot of *PSMB10* for two individuals, one carrying no rare variant in this region (control, upper tracks), and one carrying a heterozygous frameshift variant (dashed line and lower tracks), in cultured fibroblasts (top) and putamen (bottom). The frameshift variant is located on exon 4 which is included on the canonical transcript (ENST00000358514) but not on transcript ENST00000570985. On the right, the barplots show the transcript expression proportions on each tissue on average across GTEx. **(b)** Fold change of gene expression against normalized gene expression rank for *PSMB10* in fibroblasts and putamen (basal ganglia) brain tissues. *PSMB10* is an expression outlier (red) in individual 2 in fibroblast but not in putamen, consistent with the rare variant triggering nonsense-mediated decay and leading to a strong gene expression reduction in the tissue for which the exon 4-containing transcript is the major isoform.

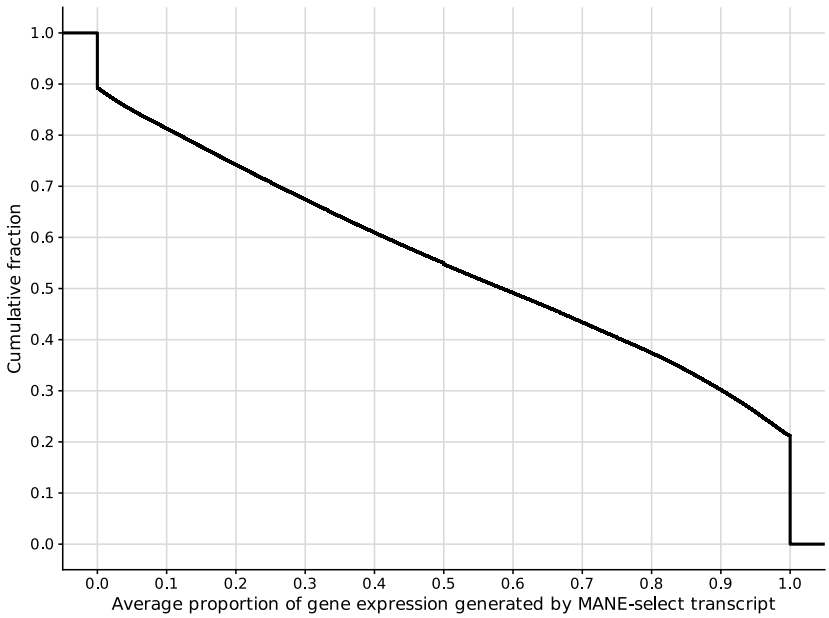

Figure 4.7: Most canonical transcript isoforms contribute only a fraction of the total gene expression. The figure shows the cumulative fraction of genes (y-axis) for which the canonical transcript (MANE-select) contributes to more than a given proportion (x-axis) to the total gene expression. The data is aggregated over all tissues and restricted to the expressed genes per tissue.

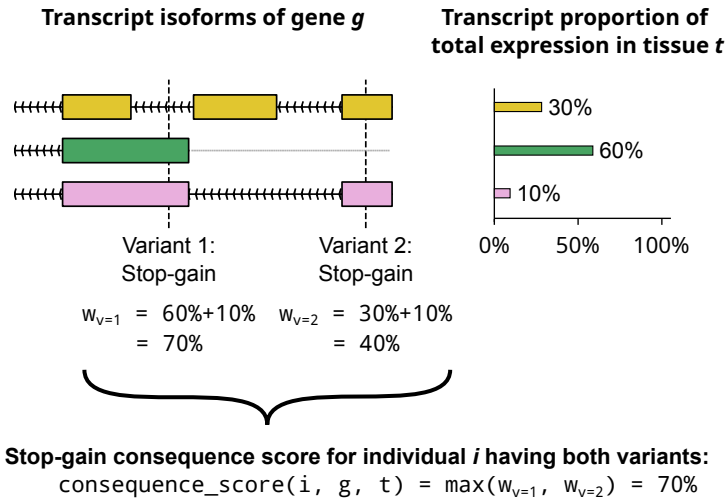

Figure 4.8: Illustration of stop-gain consequence score calculation for two variants. Variant 1 causes a stop-gain in $w_{v=1} = 70\%$ of all expressed isoforms, variant 2 in $w_{v=2} = 40\%$ of all expressed isoforms. Therefore, the stop-gain consequence score for this individual i in gene g and tissue t is 70%.

gene g, individual i, and tissue t across variants with a certain consequence v:

$$\text{consequence\_score}(i, g, t) = \max_{v \in \text{variants}(i,g)} w_v(t) \tag{4.4}$$

Figure 4.8 shows an illustration of how this consequence score is being calculated for two stop-gain variants.

Training a model using these tissue-specific weighted annotations more than tripled the precision for the highest scoring predictions and significantly increased the average precision by 55% to reach 2.7% in median across tissue types (fig. 4.9).

4.5 Incorporating the tissue-specific gene expression variability

OUTRIDER not only models the mean expression levels of genes but, similar to other statistical models for RNA-seq data, it also includes a measure of gene expression variability known as the biological coefficient of variation (see section 2.5). In the GTEx dataset, this biological coefficient of variation captures the expression variability of genes per tissue across the population.

I hypothesized that the same fold changes in expression might be considered outliers for genes with low expression variability, but not for those with high variability. Indeed, the smallest fold-change among outliers observed in the GTEx dataset decreased with the biological coefficient of variation, as shown in fig. 4.10a. Therefore, a certain reduction in gene expression might cause aberrant expression in one gene or tissue, but not in

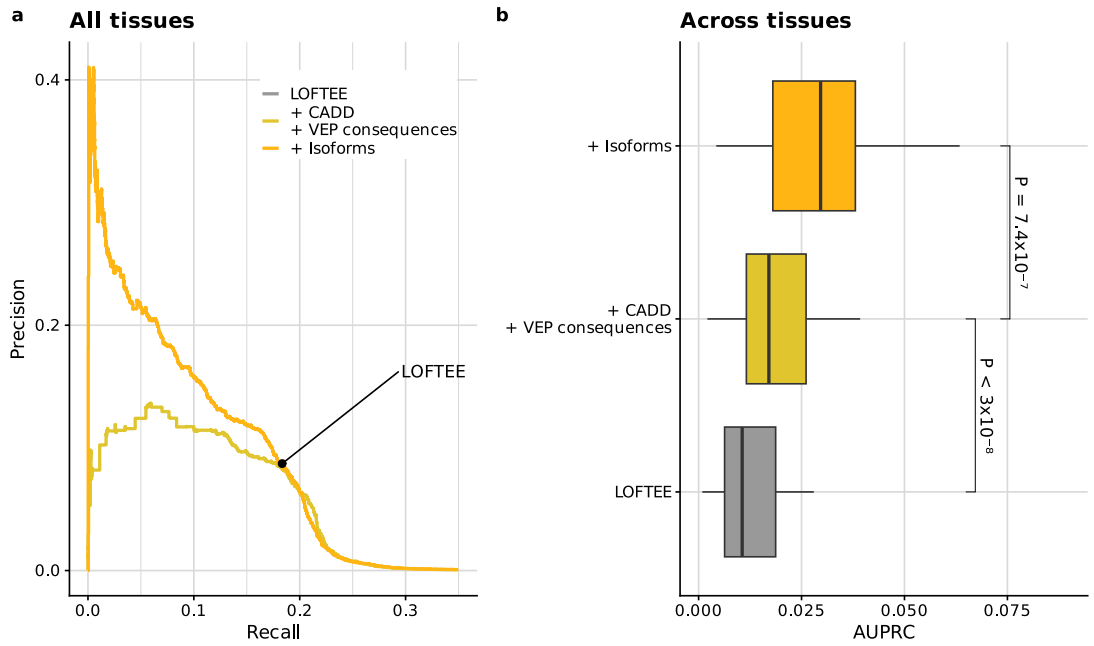

Figure 4.9: Accounting for tissue-specific isoform expression improves predictions. **(a)** Precision-recall curve for all tissues combined. LOFTEE shows up as a single point because it is a binary filter. **(b)** Distribution of average precision (AUPRC) across 26 GTEx tissue types. P-values were obtained using a paired Wilcoxon test. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

another. For example, a 30% reduction in *LTBP3* expression in the tibial artery was sufficient to be considered an outlier. In contrast, a 30% reduction in *OR2W3* expression in blood would not result in an outlier, as *OR2W3* showed a larger expression variability in blood, ranging from 10% to 230% (fig. 4.10b). *OR2W3* is one of the over 800 human olfactory receptor genes whose defect is usually benign[127]. In contrast, *LTBP3* is a gene whose dysfunction is associated with dental anomalies and short stature[29]. Therefore, it is not surprising that gene expression levels in *OR2W3* exhibit a wider physiological range than in *LTBP3*. Overall, genes with lower expression variability were more genetically constrained in the human population (i.e. harbored fewer loss-of-function variants, fig. 4.10c), in agreement with previous studies on primates[36].

To take this variability in gene expression into account, I first considered modelling expression fold-changes of variants and deriving a z-score $\hat{z}$ from the OUTRIDER-fitted negative binomial distribution by estimating the observed count $\hat{x}$ as a product of the expected read count μ and the predicted fold-change $\hat{\text{fc}}$:

$$\hat{x}_{i,g,t} = \mu_{g,t} \cdot \hat{\text{fc}}_{i,g,t} \tag{4.5}$$

$$\hat{z}_{i,g,t} = \text{CDF}^{-1}_{N(0,1)}\left(\text{CDF}_{NB}(\hat{x}_{i,g,t}|\mu_{g,t}, \theta_{g,t})\right) \tag{4.6}$$

, where $\text{CDF}^{-1}_{N(0,1)}$ is the inverse cumulative distribution function of the standard normal distribution and $\text{CDF}_{NB}(x|\mu,\theta)$ the negative-binomial cumulative distribution function. This approach assumes that a variant affects the gene expression fold-changes independently of the expression variability. In other words, a heterozygous high-impact variant that completely perturbs the expression of one copy of the gene is assumed to reduce the gene expression by a fixed fraction, e.g. 50%, regardless of the expression variability. However, when testing this assumption using variants likely triggering NMD, I noticed that fold-changes of the same class of variants were correlated with expression variability (fig. 4.11). An explanation for this correlation between fold-changes and expression variability could be that genes with low expression variability are subject to regulatory buffering mechanisms[7, 31].

Therefore, I chose a more general modeling approach by providing the biological coefficient of variation as an additional input feature to the non-linear model predicting the z-score. This model increased the performance by more than 50% to 4.0% average precision (median across tissue types, fig. 4.12). These findings demonstrate the importance of considering gene expression variability in the prediction of aberrantly expressed genes, and that the use of z-scores for predictions provides more relevant insights for variant interpretation than relying solely on fold-change.

4.6 Contribution of aberrant splicing and transcript ablations

As shown in chapter 3, tissue-specific aberrant splicing predictions from AbSplice-DNA were strongly enriched in underexpression outliers and predictive for underexpression

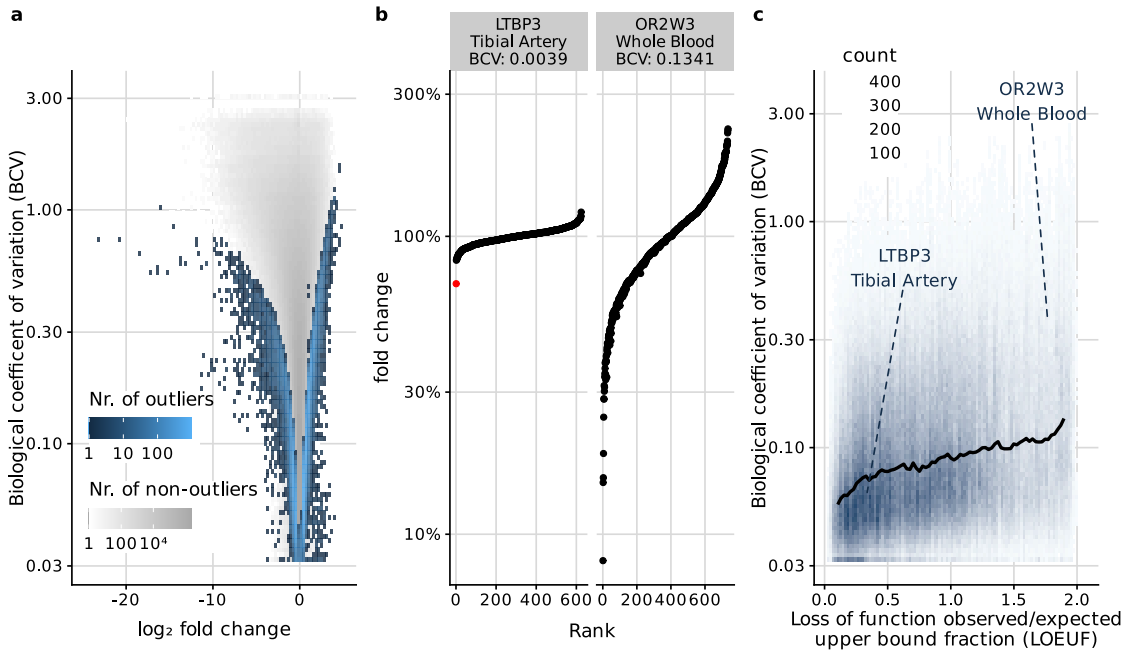

Figure 4.10: The outlier state depends on the tissue-specific biological coefficient of variation. **(a)** Biological coefficient of variation (BCV) against expression fold change across all genes and tissues. Highly variable genes require a larger fold change to be called an outlier. **(b)** Fold change of gene expression against normalized gene expression rank for *LTBP3* in tibial artery, an autosomal recessive gene whose defect can lead to dental anomalies and short stature 21, and for *OR2W3* in blood, an olfactory gene whose defect should not impair the viability of an individual 22. Expression outliers are highlighted in red. *LTBP3* is tightly regulated with a fold change range of ± 20% among non-outliers. The individual marked in red carries a heterozygous frameshift variant that associates with 30% reduction and which is detected as an outlier. In contrast, *OR2W3* shows very large variations where individuals with 30% reductions are not outliers. **(c)** BCV versus loss of function observed/expected upper bound fraction 17 (LOEUF) across all genes and tissues. Genes with a high LOEUF are more tolerant to loss of function. The black line shows a running median between LOEUF and BCV. The autosomal recessive gene *LTBP3* has a low LOEUF, denoting a low loss of function tolerance. In contrast, the olfactory gene *OR2W3* has a high LOEUF, denoting a large loss of function tolerance.

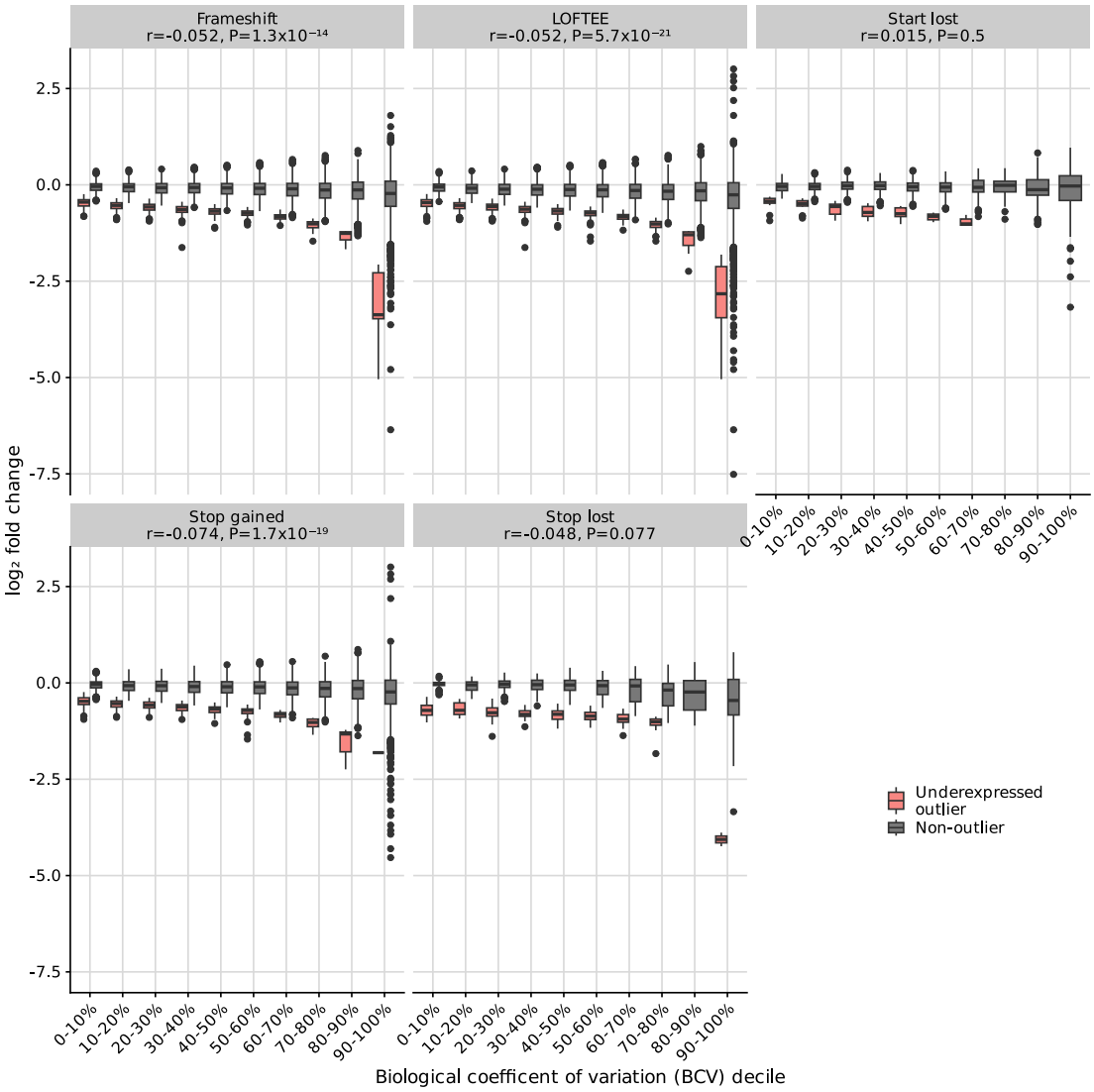

Figure 4.11: Fold-changes of the same class of variants are correlated with expression variability. This figure shows the distribution of gene expression fold changes among genes in different deciles of expression variability, given that the gene is affected by some rare variant consequence (e.g. frameshift, LOFTEE). Denoted in the facet title are Spearman's r between fold-changes and BCV as well as its significance. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. The higher the coefficient of variation, the larger the gene expression impact of the variants tends to be. If a gene has rare variants with multiple consequences, they will appear in all the corresponding panels.

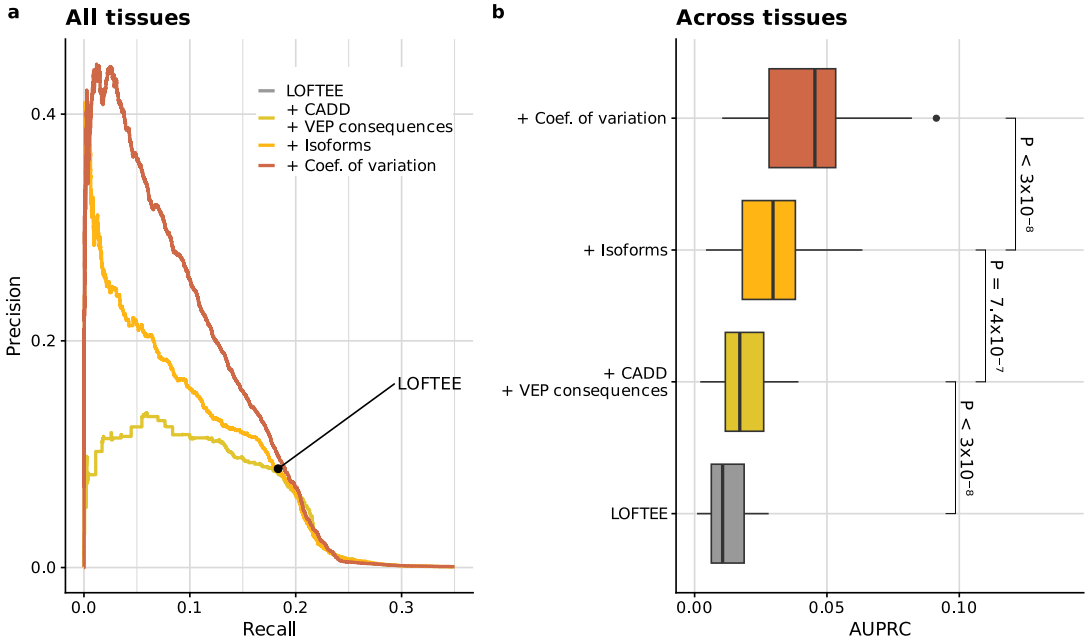

Figure 4.12: Incorporating the tissue-specific gene expression variability improves predictions. (a) Precision-recall curve for all tissues combined. LOFTEE shows up as a single point because it is a binary filter. **(b)** Distribution of average precision (AUPRC) across 26 GTEx tissue types. P-values were obtained using a paired Wilcoxon test. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

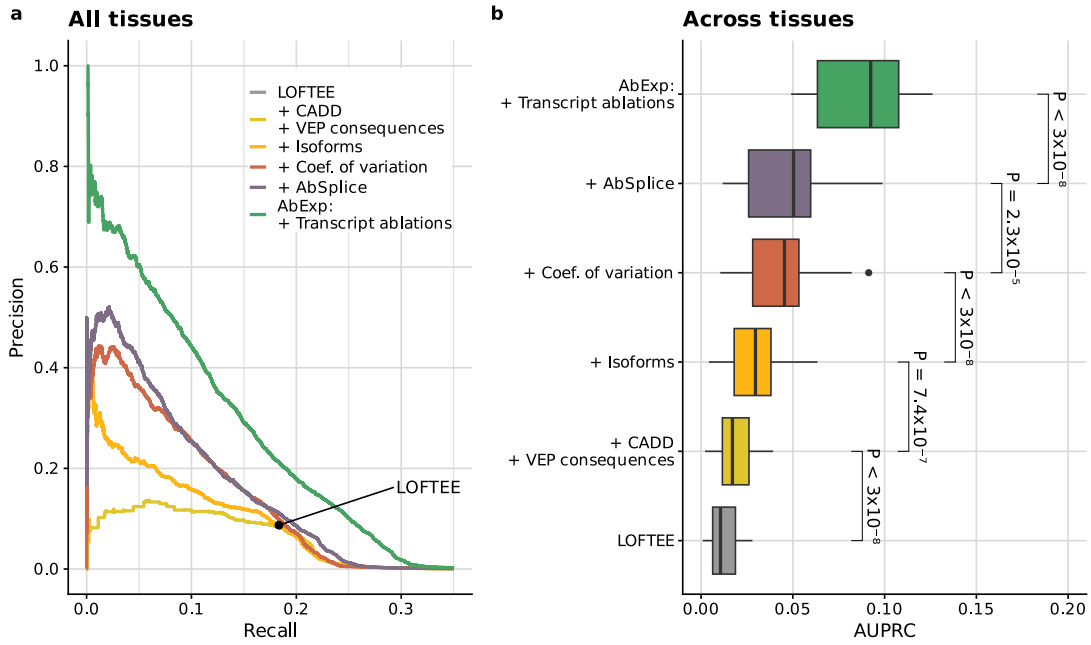

Figure 4.13: AbExp combines various variant and tissue annotations to predict aberrant gene expression and outperforms LOFTEE by about sevenfold. (a) Precision-recall curve for all tissues combined. LOFTEE shows up as a single point because it is a binary filter. **(b)** Distribution of average precision (AUPRC) across 26 GTEx tissue types. P-values were obtained using a paired Wilcoxon test. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

outliers. Therefore, I next integrated AbSplice-DNA scores as additional feature to the model. This led to a significantly improved model performance of 4.9% average precision in median across tissue types (fig. 4.13).

Another type of variants that exhibited a strong enrichment in underexpression outliers was transcript ablations derived from structural deletion variants. Including these variants into the model yielded a large gain in precision among the top-ranked predictions and increased the average precision to 9.1% in median across tissue types.

In the following, I refer to this model which integrates all features mentioned so far as AbExp. AbExp takes as input a set of variants within 5,000 bp of any annotated transcript of a protein-coding gene and returns a predicted z-score for each of the 48 tissues. As a high-confidence cutoff, I suggest a cutoff of AbExp < -3.4 corresponding to 50% precision and 8.5% recall on the benchmark data, and a low-confidence cutoff of AbExp < -1.3 corresponding to 20% precision and 19.1% recall.

4.7 AbExp performance replicates on independent datasets

Next, I evaluated whether the underexpression prediction performance of AbExp replicates on two independent datasets. The first dataset consisted of individuals suspected to be affected by a mitochondrial disorder[82] with whole-exome sequencing data paired with RNA-seq from fibroblasts (see section 2.12.3). The second dataset consisted of whole-genome and RNA sequencing measurements in amyotrophic lateral sclerosis (ALS) patients and healthy controls from the AnswerALS research project[10] (see section 2.12.2).

Structural variant calls, and thus transcript ablation calls, were not available on either dataset.

4.7.1 Outlier calling and rare variant filtering

Both the mitochondrial disease dataset and the ALS dataset were processed similar to the GTEx dataset. First, I filtered gene expression outlier calls obtained with OUTRIDER for a sufficiently large expected number of fragments ($\mu > 450$) and removed samples with more than 50 outliers. Next, I filtered variants for a genotype quality ≥ 30 and read depth ≥ 10 reads. I subsetted rare variants based on the gnomAD population with a minor allele frequency ≤ 0.001. Finally, I applied AbExp to the rare variants for outlier prediction and kept for each gene, tissue and individual the lowest AbExp score.

The mitochondrial disease dataset[149] consisted of 311 whole-exome sequencing samples paired with RNA-seq from fibroblasts. Thus, I used AbExp predictions from fibroblasts for evaluation. After filtering, this dataset contained 501 underexpression outliers across 299 samples.

For the amyotrophic lateral sclerosis (ALS) dataset, I downloaded 244 transcriptomes with matched whole-genome sequencing data from https://dataportal.answerals.org[10]. The data consisted of 205 cases diagnosed with amyotrophic lateral sclerosis and 39 control samples. RNA-seq measurements were obtained from iPSC-derived spinal motor neurons, specialized nerve cells located in the spinal cord. Therefore I used AbExp predictions for the tibial nerve, considering it the most relevant tissue type, for the evaluation. After filtering, the dataset contained 739 underexpression outliers across 244 samples. A detailed overview of how many samples, genes, etc. remained after each filtering step in both datasets can be seen in table 4.1 and table 4.2.

4.7.2 Performance evaluation

In both the mitochondrial disease and ALS datasets, AbExp significantly outperformed the baseline methods CADD and LOFTEE (fig. 4.14). On GTEx, AbExp had performed two or three times better than LOFTEE and CADD in average precision, without taking transcript ablation annotations into account. In the context of the mitochondrial disease and ALS datasets, AbExp without transcript ablation annotation not only enabled slightly better precision at the same recall compared to LOFTEE filtering but also provided a continuous score that facilitated achieving much higher precisions. Notably,

Filtering step	Individuals	Genes	Tissues	Samples	Under-expressed outliers	Non-outliers	Over-expressed outliers
Unfiltered	325	14740	1	325	430021	3923672	436807
Samples with whole genomes	311	14740	1	311	411232	3754846	418062
Keep only protein-coding genes	311	11849	1	311	342540	3000566	341933
Remove samples with many outliers	302	11849	1	302	328541	2921805	328052
Keep only genes of samples that have sufficiently large expected number of reads ($\mu > 450$)	302	11316	1	302	242648	2165069	240421

Table 4.1: Mitochondrial disease dataset filtering.

Filtering step	Individuals	Genes	Tissues	Samples	Under-expressed outliers	Non-outliers	Over-expressed outliers
Unfiltered	253	16381	1	253	1919	4141190	1284
Keep only protein-coding genes	253	12771	1	253	1608	3228306	1149
Remove samples with many outliers	244	12771	1	244	1268	3114031	825
Keep only genes of samples that have sufficiently large expected number of reads ($\mu > 450$)	244	11516	1	244	739	1484062	443

Table 4.2: ALS dataset filtering.

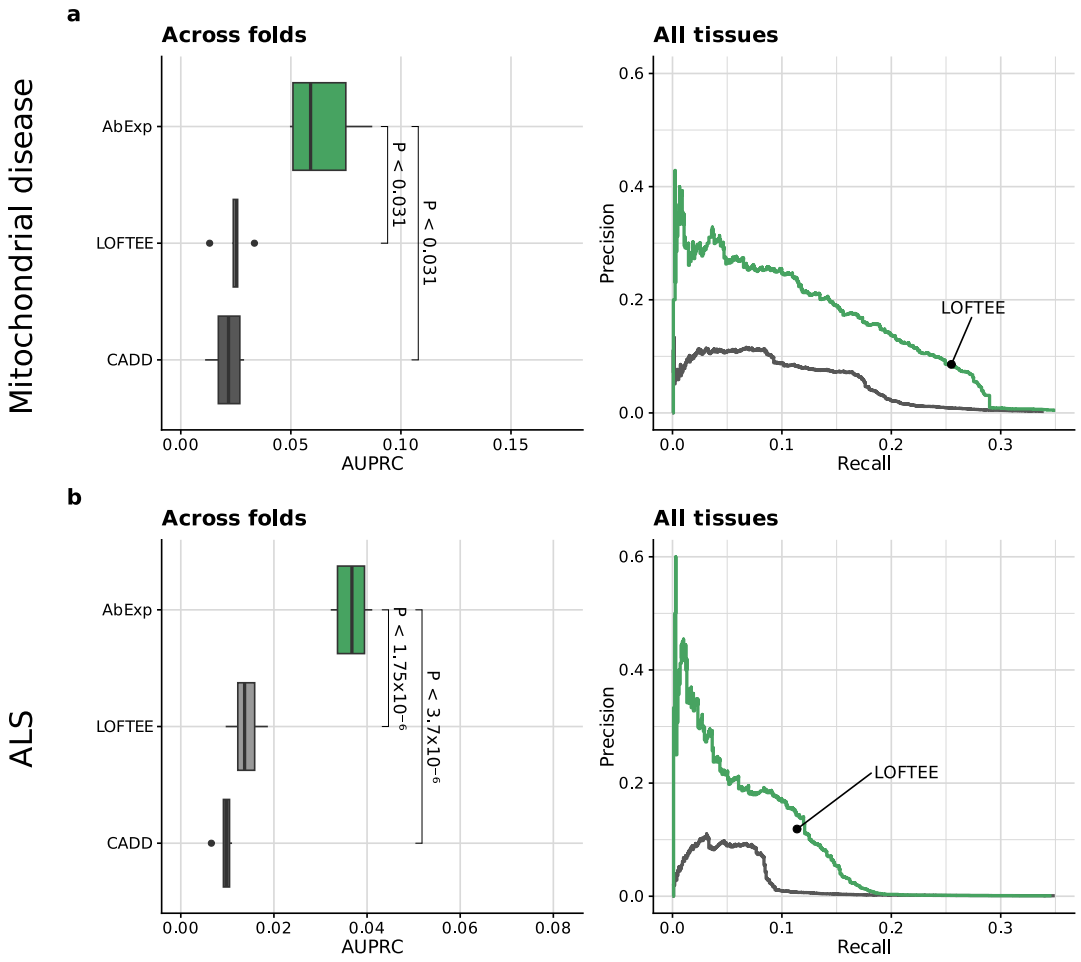

Figure 4.14: Performance of AbExp replicates on independent datasets. **(a)** Left: Distribution of average precision (AUPRC) across five cross-validation folds in the mitochondrial disease dataset. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. P-values were obtained using a paired Wilcoxon test. Right: Precision-recall curve on the whole mito-disease dataset. LOFTEE as a binary predictor is shown as a single point. **(b)** Left: Distribution of average precision (AUPRC) across five cross-validation folds in the ALS dataset. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. P-values were obtained using a paired Wilcoxon test. Right: Precision-recall curve on the whole ALS dataset. LOFTEE as a binary predictor is shown as a single point.

the recall for all methods was twice as low on the ALS dataset than in the GTEx and the mitochondrial disorder dataset. This could be caused by poorer expression outlier calls, a stronger role of epigenetic and trans-regulatory effects, or a combination of these possibilities.

4.8 Analysis of AbExp scores

4.8.1 AbExp predicts on average 1.2 high-confidence and 5.7 low-confidence underexpressed genes per individual

In diagnostics, researchers must prioritize their analysis and select the most likely causal variants for further analysis. Therefore, a large number of predicted outliers could be a disadvantage. In the GTEx benchmark dataset, I observed on average 13.6 genes underexpressed in at least one tissue (fig. 4.15). A reasonable predictor should predict a comparable number of underexpression outliers. Indeed, AbExp predicts on average 1.2 high-confidence and 5.7 low-confidence outliers which is considerably lower than the number of measured underexpressed genes (fig. 4.15).

4.8.2 AbExp predictions are tissue-specific

Tissue-specific expression outliers are outliers that are aberrantly expressed in some but not all tissues. If AbExp can predict tissue-specific underexpression outliers, it should therefore be able to predict underexpression outliers in fewer tissues than the gene is expressed in. On average, a gene is expressed in 27.5 tissues (fig. 4.16a). AbExp predicts, on average, 17.0 underexpressed tissues per gene with high confidence and 20.9 underexpressed tissues per gene with low confidence, provided that the gene is predicted as underexpressed in at least one tissue (fig. 4.16a). Notably, 15% of all genes predicted to be underexpressed exhibit underexpression across all tissues with high confidence, while 29% demonstrate this pattern with low confidence (fig. 4.16b), although these percentages might be slightly skewed by genes that are expressed only in a single tissue (fig. S1). Therefore, the predictions of AbExp are indeed tissue-specific.

4.8.3 25-45% of AbExp high-confidence predictions can not be explained with LOFTEE

For a more holistic overview of what type of variants AbExp considers to be important, I predicted AbExp scores for all rare GTEx variants and selected for all gene, individual, and tissue combinations the variants with the lowest AbExp score. Further, I removed predictions of non-expressed genes (genes with less than 1 FPKM in 95% or more of the samples in a tissue, see section 3.2.1).

Based on these predictions, I examined the fraction of variant types among AbExp-predicted high-impact variants in GTEx (fig. 4.17). About 25% of predicted low-confidence underexpression outliers are LOFTEE-negative variants and, thus, can be explained

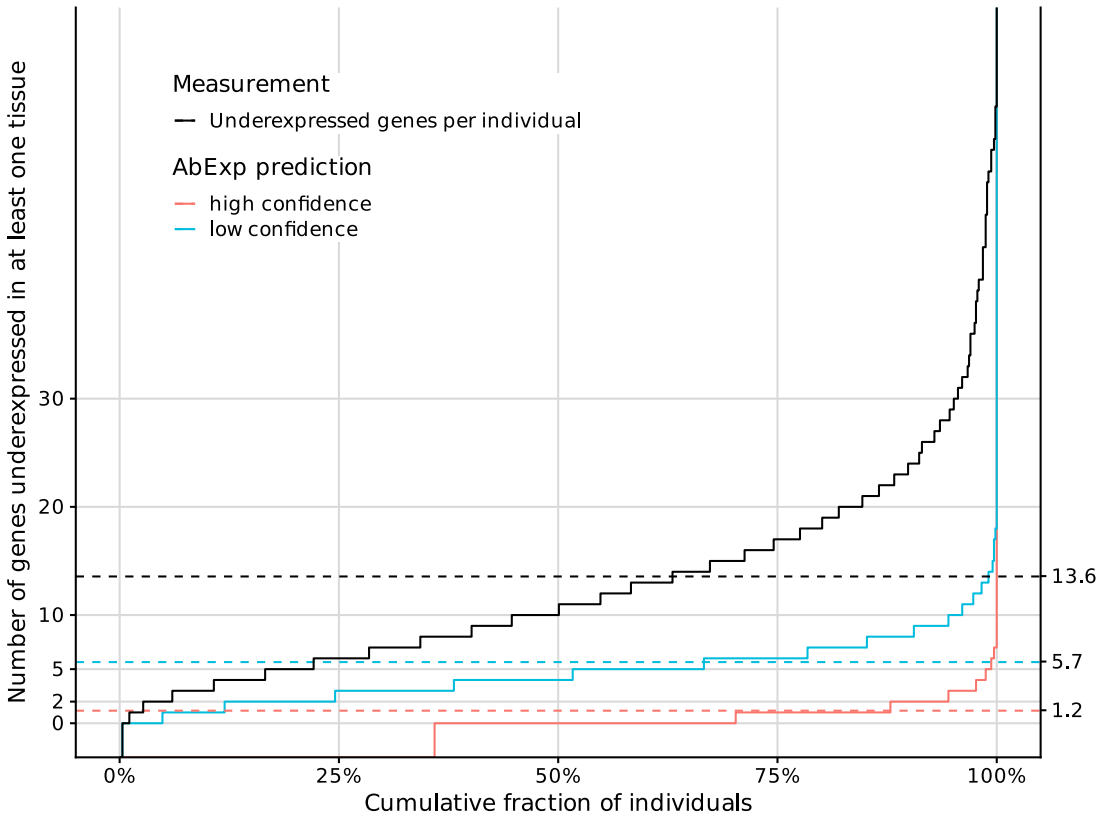

Figure 4.15: AbExp predicts on average per individual 1.2 low-confidence and 5.7 high-confidence genes to be underexpressed in at least one tissue. Fraction of individuals (x-axis) having at most a given number of genes underexpressed in at least one tissue (y-axis). The black curve shows the observed underexpressed genes per individual as identified with OUTRIDER, and the red and blue curves the number of underexpressed genes per individual as predicted by AbExp with different cutoffs. Horizontal dashed lines denote the mean of the equally colored curves.

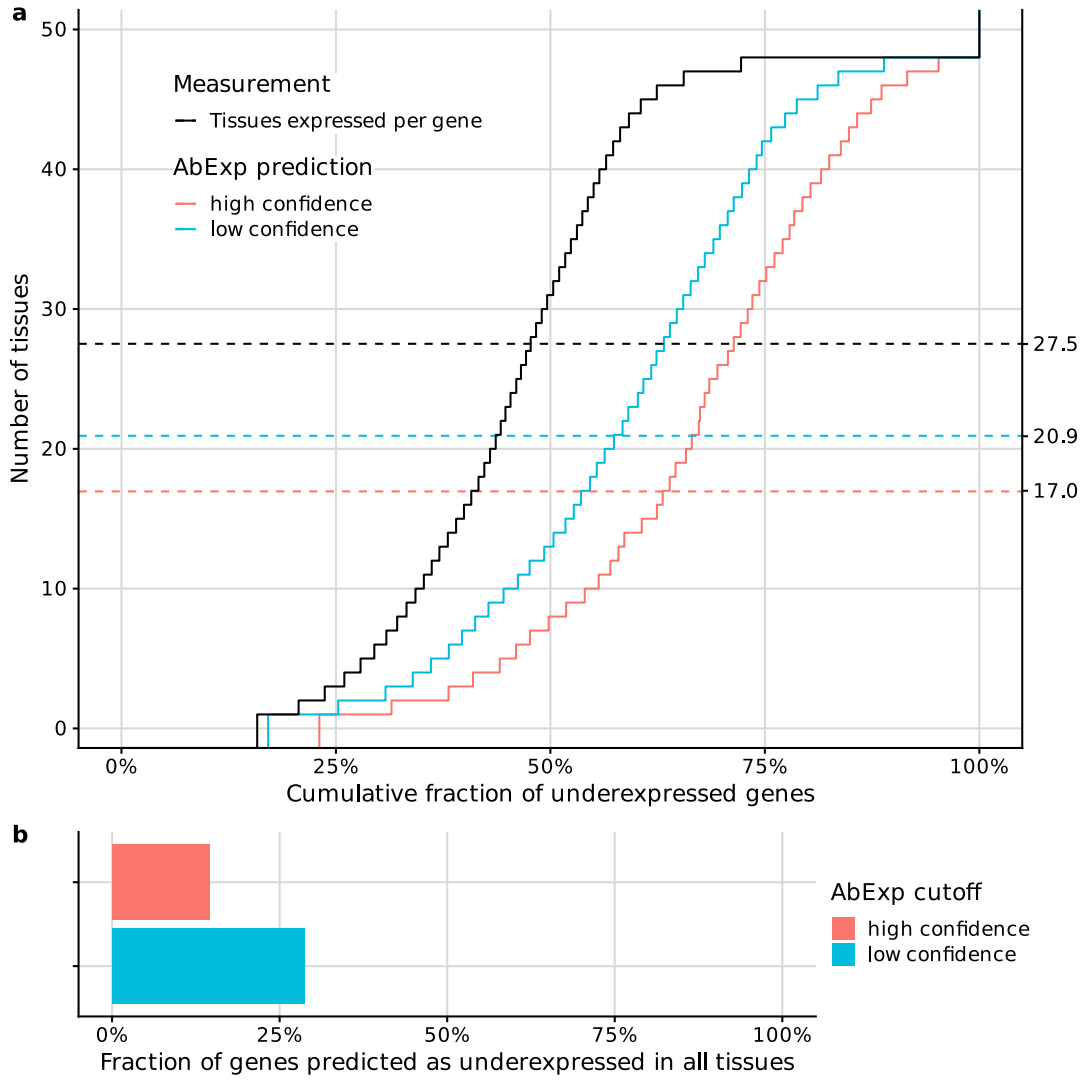

Figure 4.16: AbExp predictions are tissue-specific. (a) Fraction of underexpressed genes (x-axis) which are underexpressed in at most a given number of tissues (y-axis). The red and blue curves show the number of underexpressed genes per individual as predicted by AbExp with different cutoffs. The black curve shows the number of tissues expressed per gene. Horizontal dashed lines denote the mean of the equally colored curves. **(b)** Fraction of genes that AbExp predicts to be underexpressed in all tissues at different cutoffs.

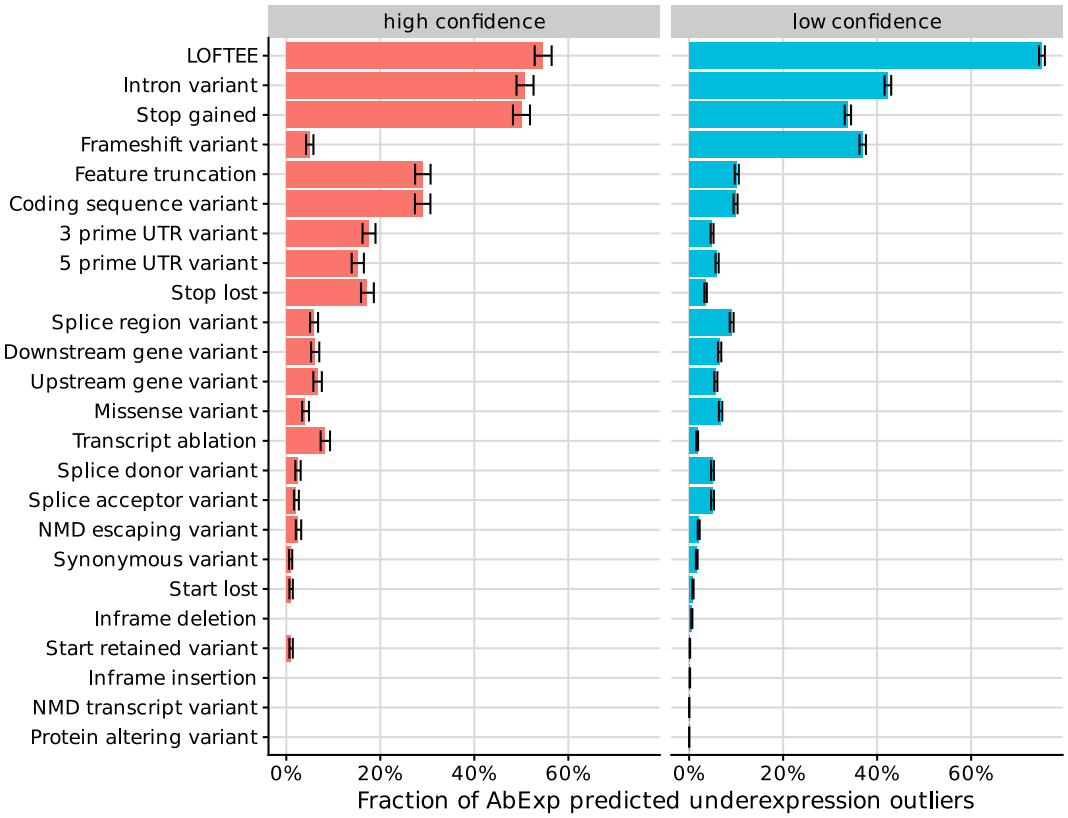

Figure 4.17: Fraction of variant types among AbExp high-impact predictions.

with AbExp but not with LOFTEE. Interestingly, this fraction is higher among high-confidence predictions, with 45% of these being LOFTEE-negative variants. Also, other variant types that LOFTEE does not consider, such as UTR variants, stop-loss variants, and transcript ablations, are more prevalent in high-confidence predictions than in low-confidence predictions.

4.8.4 AbExp correlation with measured expression varies

Upon comparing predicted and observed z-scores across various types of variants, I discovered that the Pearson correlation between them varies based on the type of variant (fig. 4.18). Notably, stop-loss variants and transcript ablations exhibit a high correlation, whereas missense variants, upstream gene variants, and downstream gene variants show minimal correlation ($R^2 < 0.005$), suggesting poor prediction accuracy for these types of variants.

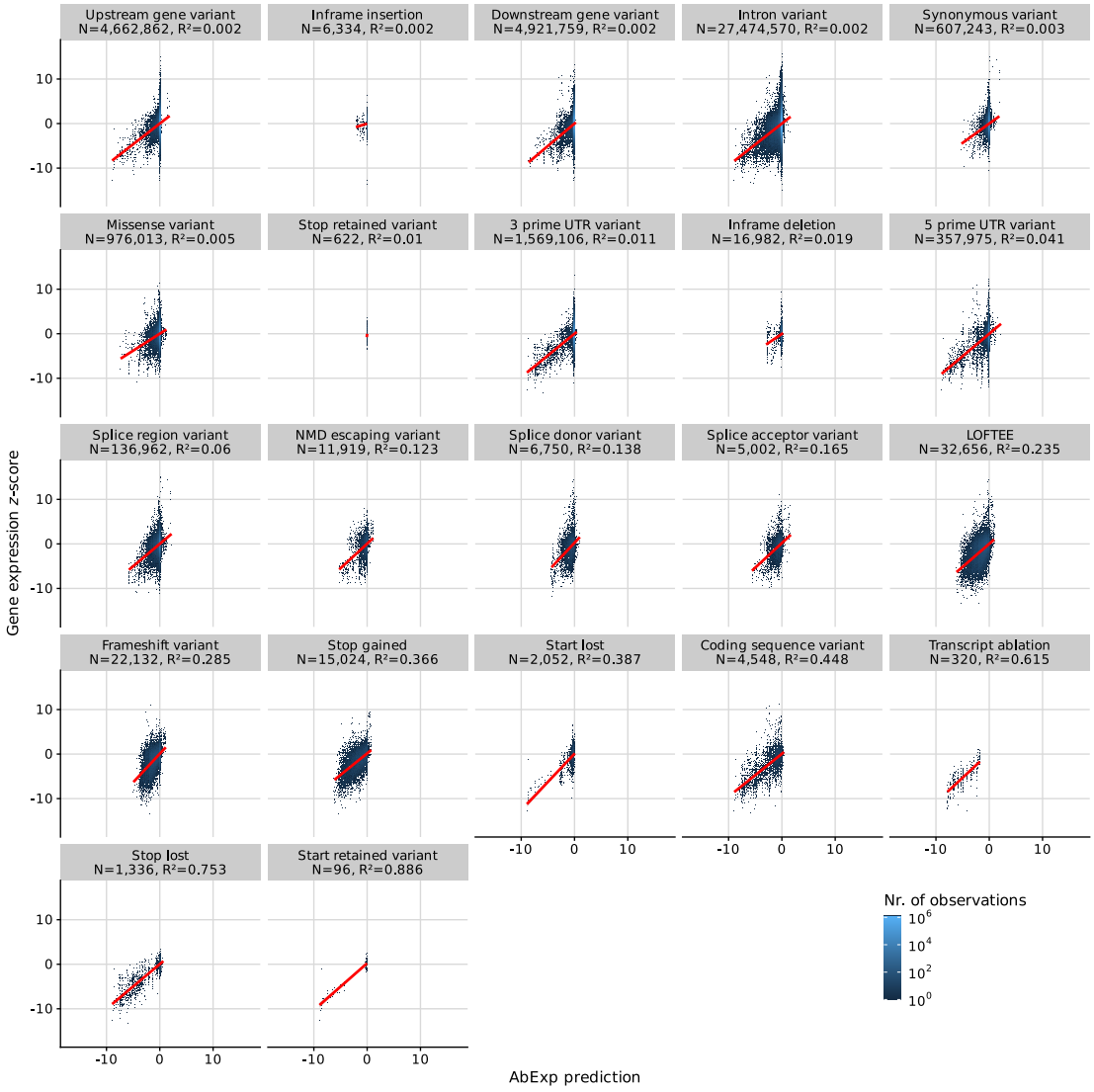

Figure 4.18: Gene expression z-score versus AbExp-DNA predictions for various types of variants. The red lines show a linear regression between predicted and observed z-scores for the corresponding variant category. N denotes the number of variants of a certain type, R^2 denotes the Pearson correlation among these.

4.8.5 AbExp predicts pathogenic variants with high precision

An important application of AbExp is the identification of high-impact variants with significant pathogenicity. To assess the efficacy of AbExp in distinguishing pathogenic from benign variants, I annotated known pathogenic, likely benign, and benign variants from the ClinVar database with both AbExp and CADD (see section 2.12.5). Importantly, I used the official CADD pipeline to annotate all variants, including those with missing pre-computed scores. Variants classified as both likely benign and benign were retained solely as likely benign within the dataset. Furthermore, for each variant, I retained only the minimum predicted AbExp score across all tissues.

In median, AbExp predicts pathogenic variants to cause gene underexpression by -1.3 standard deviations relative to the average population (fig. 4.19a). AbExp scores of likely benign and benign scores, conversely, are close to zero in the median. CADD scores are in median 8-17 times higher in pathogenic variants than in likely benign and benign variants. In comparison, AbExp scores are in median about 173-266 times higher in pathogenic variants than in likely benign and benign ones. Notably, AbExp annotated about 7% more variants than CADD.

Furthermore, employing a high-confidence threshold, AbExp successfully recalls 16.2% of all pathogenic variants with 99.6% precision. With a low-confidence threshold, AbExp achieves a recall rate of 49.7% for all pathogenic variants with 98.1% precision (fig. 4.19b). At both recall rates, CADD is less precise than AbExp, with reaching only 99.0% precision at 16.2% recall and 96.1% precision at 49.7% recall. This result emphasizes the ability of AbExp to discriminate pathogenic from benign variants with high precision. Nonetheless, CADD achieves a slightly higher average precision of 84.8% overall, compared to AbExp reaching 81.5% average precision.

4.9 The AbExp variant effect prediction pipeline

To simplify the application of AbExp, I developed a software pipeline to calculate AbExp predictions which can be found at https://github.com/gagneurlab/abexp. The input to this pipeline is (a set of) VCF files, a reference genome, and gene annotations. After setup and the configuration of file paths, the pipeline will annotate all variants with AbExp scores and return a table with the following columns:

- `chrom`: chromosome of the variant
- `start`: start position of the variant (0-based)
- `end`: end position of the variant (1-based)
- `ref`: reference allele
- `alt`: alternate allele

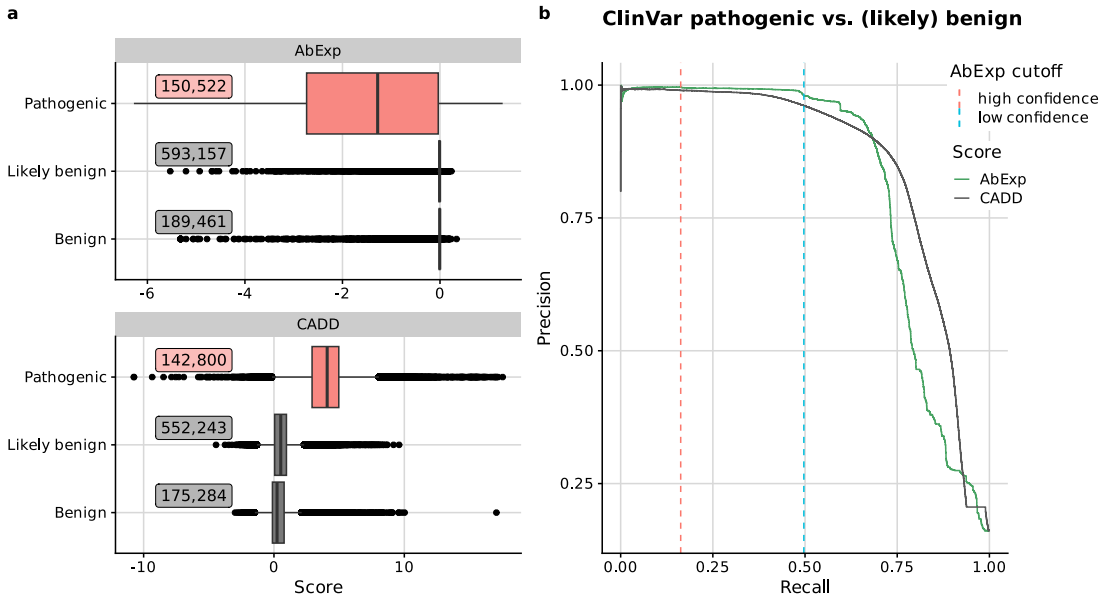

Figure 4.19: AbExp predicts pathogenic variants with high precision. **(a)** Distribution of CADD scores and minimum AbExp scores across tissues for pathogenic, likely benign, and benign variants from the ClinVar database. Box label, sample size; center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. **(b)** Precision-recall curve for distinguishing pathogenic from likely benign and benign variants using AbExp and CADD in ClinVar variants. Dashed vertical lines show the high-confidence and low-confidence thresholds of AbExp.

- `gene`: the gene affected by the variant
- `tissue`: GTEx tissue, e.g. "Artery - Tibial"
- `tissue_type`, GTEx tissue type, e.g. "Blood Vessel"
- `abexp_v1.0`: The predicted AbExp score
- a set of features used to predict the AbExp score

Further information on setup, configuration, and application can be found in the `README.md` file of the pipeline repository.

4.10 Integrating AbExp with gene expression measurements from clinically accessible tissues

In rare disease diagnostics, RNA sequencing is gaining popularity as a complementary assay to genome or exome sequencing, as it enables the direct measurement of aberrant gene regulation in a tissue of interest[24, 47, 82, 100, 110, 149]. However, many rare disorders are thought to originate from tissues that are difficult to access, such as heart or brain. Obtaining samples from these tissues can be challenging due to the highly invasive nature of the sampling process. A less invasive approach is to investigate aberrant gene regulation in clinically accessible tissues (CATs) such as skin fibroblasts or blood, as these tissues have been shown to share a substantial fraction of expressed genes with non-CATs and are therefore likely to capture aberrant expression occurring in non-CATs[24, 47, 82].

In the event that gene expression measurements from clinically accessible tissues, specifically skin-derived fibroblasts and blood, are available, I investigated their potential informativeness for predicting expression outliers. Although these measurements may not be directly transferrable to other tissues, there exists a certain degree of correlation in gene expression across tissues (fig. 4.20), potentially allowing to explain gene expression outliers in other non-accessible tissues.

To evaluate the predictive value of gene expression measurements from blood and fibroblasts, I removed the corresponding CAT along with related tissues from the predicted tissues. Specifically, when using fibroblasts as CAT, I excluded the non-sun-exposed suprapubic skin, sun-exposed lower leg skin, and cultured fibroblasts from the predicted tissues. When using whole blood as CAT, I excluded whole blood and EBV-transformed lymphocytes.

First, I used the OUTRIDER z-score in the CAT to rank underexpression outliers in non-CATs. When skin fibroblasts were used as CAT, it resulted in a notably higher AUPRC of 17.7% (median across tissue types), which is a significant improvement compared to the 8.8% achieved by the genome-based predictor AbExp (fig. 4.21a). When using whole blood as CAT (9.1% median AUPRC), there was no performance improvement compared to AbExp (9.6% median AUPRC, fig. 4.21b).

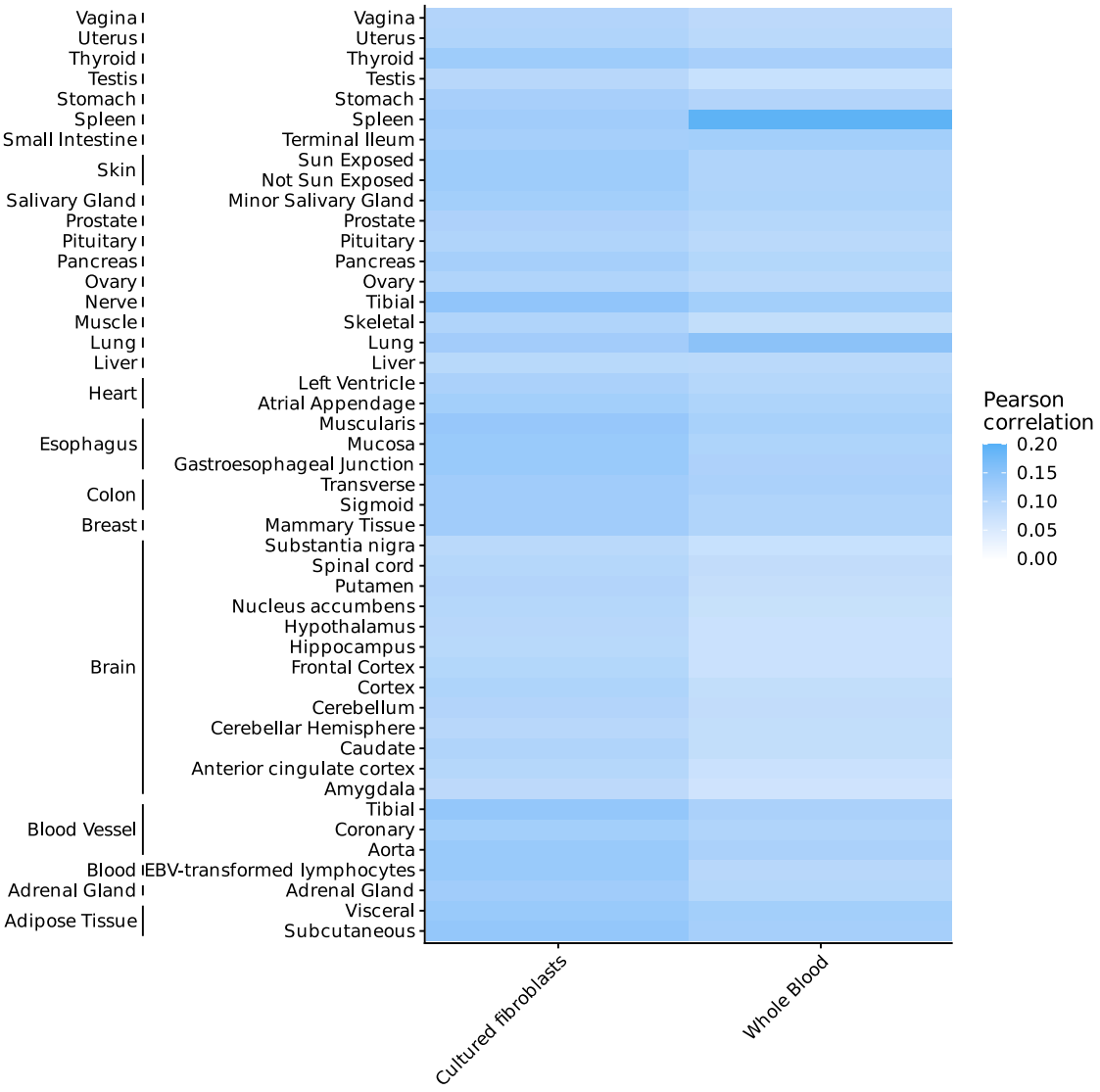

Figure 4.20: Gene expression correlates between clinically accessible tissues and non-accessible tissues. Heatmap shows the Pearson correlation of OUTRIDER z-scores between clinically accessible tissues and other tissues in GTEx.

Next, I trained a model that integrates AbExp with gene expression in a CAT to predict underexpression outliers. Contrary to the DNA-based models, I used a logistic regression (section 2.7.1) for this model. The input for this regression included:

- a binary variable indicating whether the gene is expressed in the CAT
- the OUTRIDER z-score of the gene in the CAT
- the AbExp prediction for the target tissue
- all three interaction terms between those three variables.

I used the same cross-validation scheme as for the other DNA-based models.

By utilizing RNA-seq data from skin fibroblasts to predict aberrant underexpression across all other tissues, the model achieved a median AUPRC of 19.5% across tissues (fig. 4.21a). Using whole blood as CAT results in a slightly lower performance than with using fibroblasts, reaching a median AUPRC of 16.1%. This is in line with previous studies based on shared expressed genes[2, 149] and our work on predicting aberrant splicing[144], where fibroblasts proved to be more informative than whole blood as fibroblasts express more genes.

In summary, these results demonstrate that integrating RNA-seq data from fibroblasts with genomic variant annotations from AbExp significantly improves the model's performance, doubling the average precision compared to using genomic variants alone.

4.11 Summary

In this chapter, I introduced AbExp, a DNA-based model designed to predict aberrant gene underexpression across human tissues. By integrating expression variability with the effects of variants on isoforms and aberrant splicing in a tissue-specific manner, AbExp achieved an average precision of 9.1% in median across tissue types, outperforming existing variant annotation tools between 6-fold and 18-fold. AbExp predicted on average 1.2 high-confidence and 5.7 low-confidence underexpressed genes per individual. The performance of AbExp was validated through replication in two independent datasets comprising patients with mitochondrial disease and amyotrophic lateral sclerosis, demonstrating the model's robustness and reliability. Also, AbExp scores distinguished benign and pathogenic variants with high precision. Further integration of expression measurements from clinically accessible tissues led to another two-fold improvement.

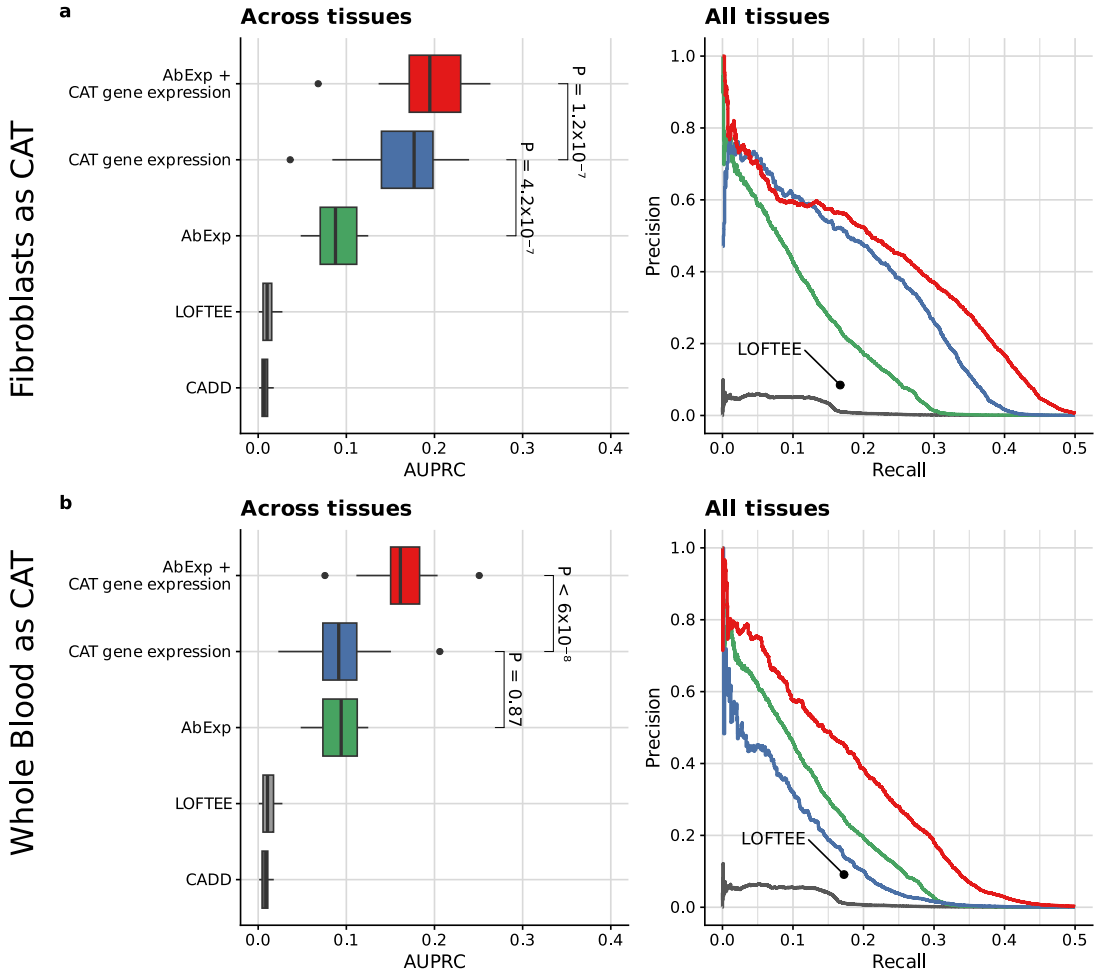

Figure 4.21: Combining RNA-seq measurements from clinically accessible tissues with AbExp improves the prediction performance. (a) Left: Distribution per predictor (rows) of average precision (AUPRC) across 25 tissue types excluding skin tissues. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. P-values were obtained using the paired Wilcoxon test. The "CAT gene expression" predictor ranks genes according to their OUTRIDER z-score in fibroblasts RNA-seq data. Right: Precision-recall curve aggregated across the same GTEx tissues as in the left panel. LOFTEE as a binary predictor is shown as a single point. **(b)** as in (a) using Whole blood as CAT and all other tissues as non-CAT.

5 Improving rare variant association testing and phenotype prediction with AbExp

5.1 Motivation

With the rise of large exome and genome sequencing biobanks, recent rare variant association testing (RVAT, see section 1.2) studies have identified gene-trait associations by leveraging the occurrence of likely high-impact (e.g. LOFTEE-positive) variants within genes, helping to pinpoint causal genes for traits[69, 145] and enabling improved phenotype predictions, particularly among individuals showing extreme phenotypes[43]. Importantly, rare expression outlier associated variants identified by integrative genomics and transcriptomics studies (section 1.3.3) have been shown to be predictive of strong effects on phenotypic traits[132]. After having established AbExp, the question arises: Can predicted aberrant expression of genes be used to predict human phenotypes?

In this chapter, I will demonstrate how AbExp and LOFTEE can be used in rare variant gene association testing and phenotype prediction (fig. 5.1). In short, I first applied AbExp and LOFTEE to whole-exome sequencing data of 200,593 individuals from the UK Biobank. The UK Biobank is a large-scale biomedical database that includes genetic, clinical, and lifestyle information from about 500,000 participants in the United Kingdom (see section 2.12.4). I then performed rare variant association testing on 40 blood traits to identify significant gene-trait associations with both LOFTEE and AbExp. Finally, I developed improved phenotype prediction models leveraging the scores from trait-associated genes on a held-out test set. Figure 5.2 shows an illustration of the data splitting.

5.2 Rare variant association testing

I used linear regression (section 2.7.1) on 40 different blood traits, including high-density lipoprotein cholesterol, glucose, and urate levels, as a common framework for rare variant association testing. Association between a trait and a gene was tested with a likelihood ratio test (section 2.8) between a restricted linear regression model containing only covariates and a full model with additional burden scores of the gene (fig. 5.3). In all regression models, I accounted for the effects of common genetic variation by including sex, age, age^2, age times sex, age^2 times sex, the first 20 genetic principal components,

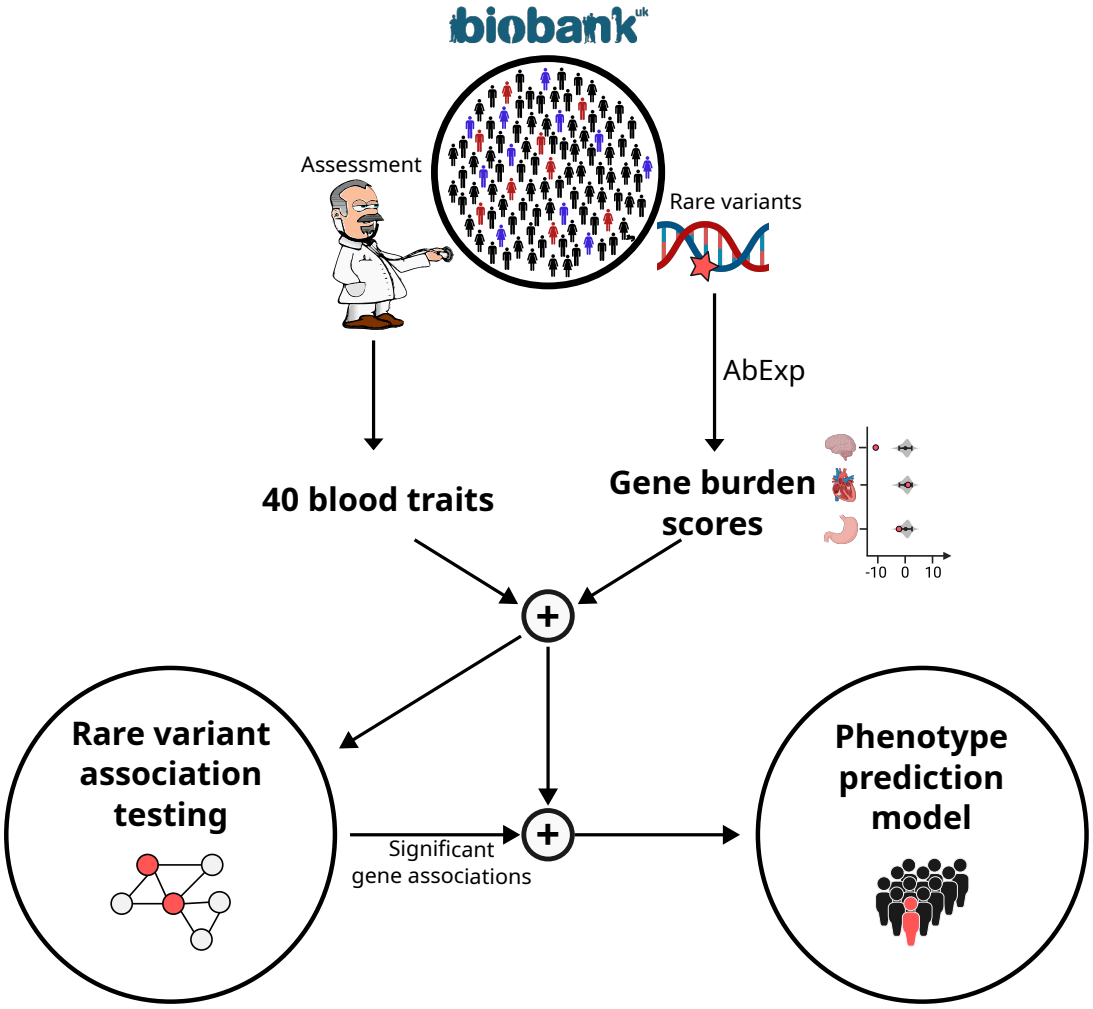

Figure 5.1: Rare variant association testing and phenotype prediction with AbExp. First, AbExp was applied to whole-exome sequencing data of 200,593 caucasian unrelated individuals in the UK Biobank. Next, rare variant association testing was performed on 40 blood traits to identify significant gene-trait associations in two-thirds of the dataset. Finally, improved phenotype prediction models were built by leveraging scores from trait-associated genes on the remaining third of the dataset.

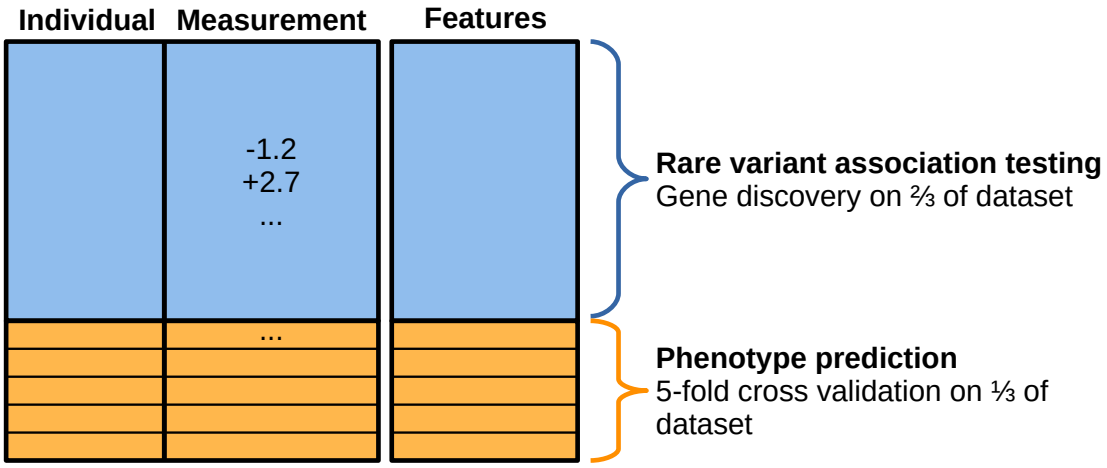

Figure 5.2: Illustration of data splitting for rare variant association testing and phenotype prediction.

and polygenic risk scores predicting the trait as covariates in the regression models. To control for common variants near each gene, I further included lead trait-associating variants within 250,000 bp of the tested gene as covariates.

To compare against a realistic baseline, I first performed RVAT using as a burden score the number of rare LOFTEE putative loss-of-function variants in each gene, similar to the Genebass study[69]. I then performed RVAT leveraging tissue-specific AbExp predictions by considering the lowest AbExp-predicted z-score across all rare variants for each of the 48 tissues. Additionally, to investigate the added value of tissue-specific AbExp predictions, I compared this tissue-specific RVAT with aggregated RVATs using the minimum and the median of the 48 values as burden scores (table 5.1).

5.2.1 UK Biobank genome and phenotype data

For this study, I used data from 200,593 caucasian unrelated individuals in the UK Biobank (fields 22006 and 22011). All of these had genotypes available from both exome-sequencing and microarrays as well as blood and urine measurements. While exome sequencing reveals genetic variation in coding regions of the genome, the UK Biobank consortium used microarray genotyping to specifically test for single nucleotide variations at roughly 800,000 positions across the whole genome. To account for population structure, I used genetic principal components derived from microarray genotyping data (field 22009). As done in the Genebass study[69], I normalized all trait values using inverse rank normal transformation. A detailed list of the used blood phenotypes can be found in table 5.2.

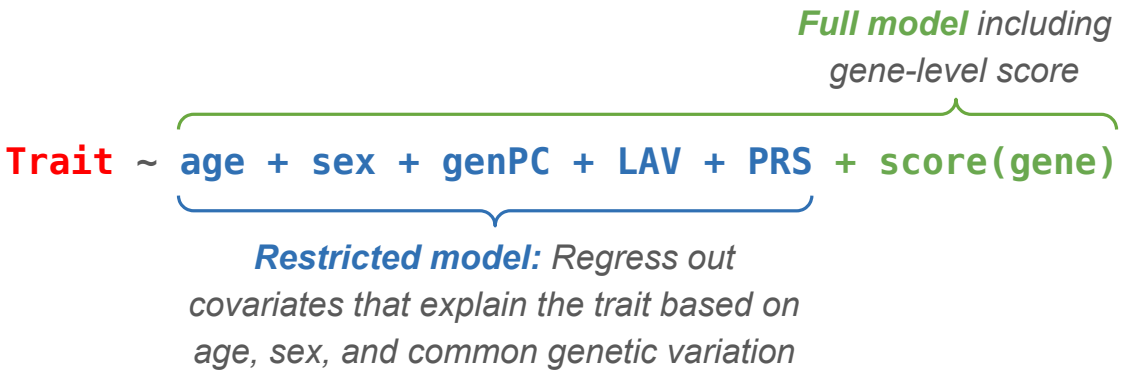

Figure 5.3: Accounting for common variation in RVAT. Each gene-trait association test is a likelihood ratio test between a restricted model containing only covariates and a full model with additional burden scores of the gene. Covariates comprise sex, age, the first 20 genetic principal components, lead trait-associating variants within ± 250,000 bp of the gene, and a polygenic risk score predicting the trait.

Model	Predictor variables	Nr. of variables
LOFTEE	Number of LOFTEE variants in the gene of interest	1
AbExp all tissues	The set of minimal AbExp scores across all rare variants in the gene (0 in absence of any rare variant) for each tissue	48
Minimum AbExp	The minimum AbExp score across all rare variants in the gene and across all tissues, 0 in absence of any rare variant	1
Median AbExp	The median AbExp score across all rare variants in the gene and across all tissues, 0 in absence of any rare variant	1

Table 5.1: Rare variant association test models.

Trait	PGS catalog ID	UK Biobank field code
Alanine aminotransferase	PGS001940	30620
Albumin	PGS001886	30600
Alkaline phosphatase	PGS001939	30610
Apolipoprotein A	PGS001888	30630
Apolipoprotein B	PGS001889	30640
Aspartate aminotransferase	PGS001941	30650
C reactive protein	PGS001946	30710
Calcium	PGS001893	30680
Cholesterol	PGS001895	30690
Creatinine	PGS001945	30700
Cystatin C	PGS001947	30720
Direct bilirubin	PGS001942	30660
Eosinophill count	PGS001172	30150
Erythrocyte distribution width	PGS001908	30070
Gamma glutamyltransferase	PGS001964	30730
Glucose	PGS001952	30740
Glycated haemoglobin (HbA1c)	PGS001953	30750
Haematocrit percentage	PGS001925	30030
HDL cholesterol	PGS001954	30760
IGF1	PGS001960	30770
LDL direct	PGS001933	30780
Leukocyte count	PGS001962	30000
Lipoprotein A	PGS001963	30790
Lymphocyte percentage	PGS001986	30180
Mean corpuscular haemoglobin	PGS001989	30050
Mean corpuscular volume	PGS001990	30040
Mean reticulocyte volume	PGS000987	30260
Mean sphered cell volume	PGS002008	30270
Monocyte count	PGS001968	30130
Neutrophill percentage	PGS001997	30200
Phosphate	PGS001998	30810
Platelet count	PGS001973	30080
Reticulocyte count	PGS001528	30250
SHBG	PGS001977	30830
Testosterone	PGS001988, PGS001914	30850
Thrombocyte volume	PGS001971	30100
Total bilirubin	PGS001942	30840
Triglycerides	PGS001979	30870
Urate	PGS002010	30880
Vitamin D	PGS001982	30890

Table 5.2: List of blood traits and corresponding PGS catalog IDs of polygenic risk scores.

5.2.2 Identification of lead trait-associated common variants

As a source of trait-associated common variants I used data from the Pan-ancestry genetic analysis of the UK Biobank (Pan-UKBB)[114]. To identify independent lead variants for every trait, Jonas Lindner used plink v1.9[18, 128] to clump variants with an association p-value ≤ 0.0001 in 250 kbp windows with an LD threshold of $r^2 < 0.5$ and then subsetted the imputed genotypes for these lead variants in a 250 kbp window around each gene.

5.2.3 Application of polygenic risk scores

I selected polygenic risk scores from the PGS catalog database[84] and Jonas Lindner applied these to the imputed genotypes using plink v2.0[18, 129]. If available, I selected scores from a study by Privé et al.[119], otherwise from a study by Tanigawa et al.[139]. A list of PGS catalog IDs used for each trait can be found in table 5.2.

5.2.4 Variant filtering and annotation

Similar to the GTEx dataset, I filtered variants for a genotype quality ≥ 30 and read depth ≥ 10 reads and subsetted rare variants based on the gnomAD population with a minor allele frequency ≤ 0.001. I then annotated all rare variants using Ensembl VEP[104] v108 with the LOFTEE plugin[70] and AbExp.

5.2.5 P-value calculation and calibration

I computed P-values for gene-trait associations using two-thirds of the dataset and applied Bonferroni correction to adjust for multiple testing. I considered gene-trait associations statistically significant if their Bonferroni-adjusted P-value was less than or equal to 0.05. As the polygenic risk scores and lead variants were based on the UK Biobank dataset, there is a possible data leakage that may have led to model overfitting. However, this does not affect the comparison between restricted and full models since all compared models always include the same set of features as covariates. P-value calibration was tested by random shuffling of the phenotypes without replacement. I found that all P-values were calibrated as can be seen in the quantile-quantile plots in fig. 5.4.

In total, I identified 28% more gene-trait associations using AbExp predictions in 48 tissues than with the LOFTEE-based model (fig. 5.5), demonstrating that AbExp can improve RVAT-based gene discovery. Further, association testing using tissue-specific predictions outperformed aggregated forms of the AbExp score in most cases by identifying a greater number of gene-trait associations (fig. 5.5b).

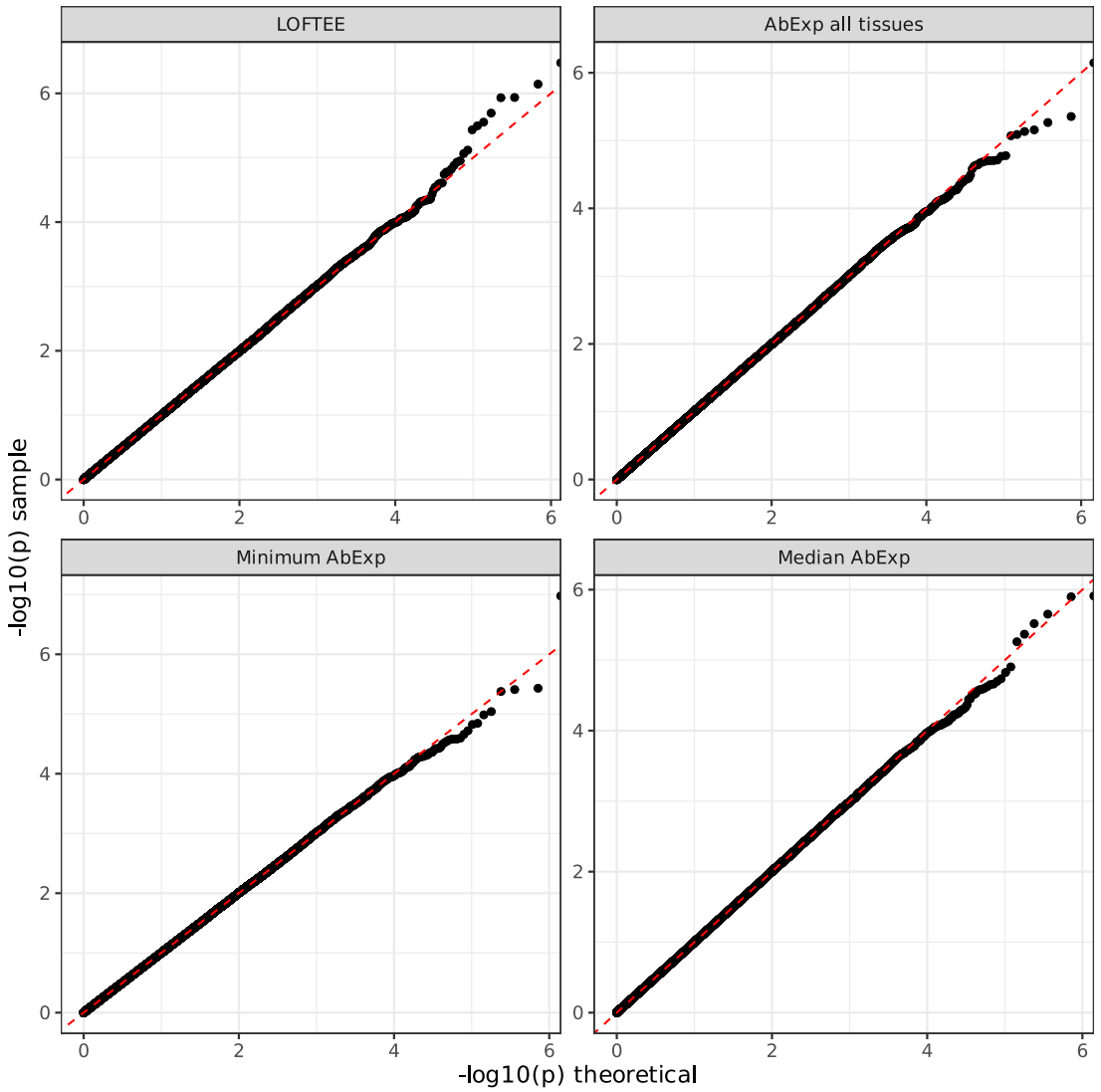

Figure 5.4: P-values of all models are calibrated. The figure shows a quantile-quantile plot of p-values on phenotype-permutated data against a random uniform distribution across all traits, facetted by the model. Models are calibrated when the data aligns closely with the diagonal (red dashed line). Across all the four models and 40 traits, I found only two significant associations on these permuted data, one for AbExp in the "Total bilirubin" trait and one for LOFTEE in the "IGF1" trait.

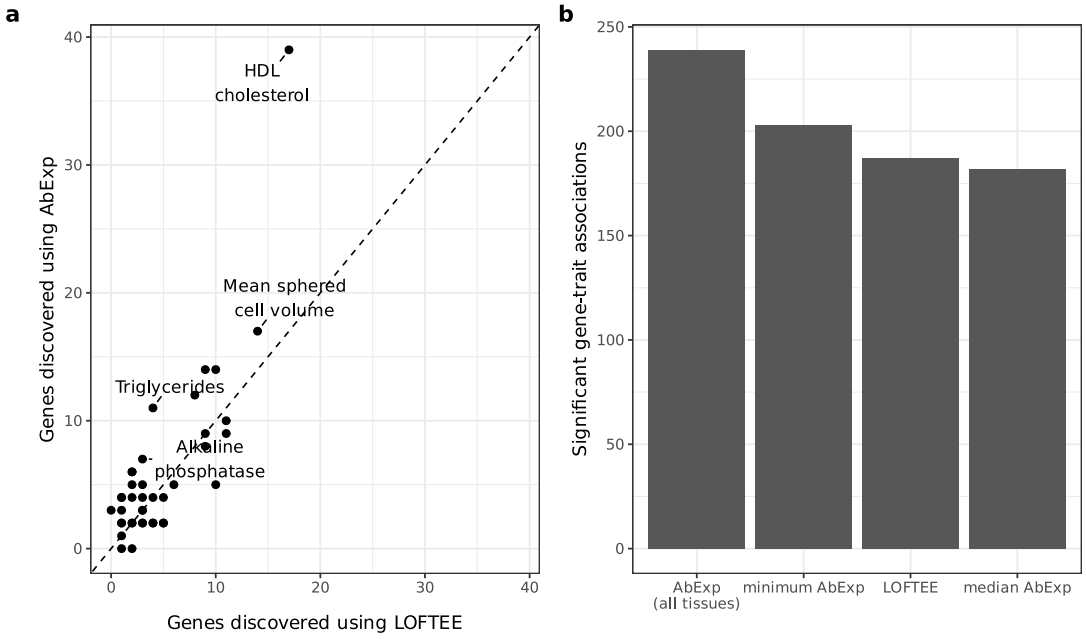

Figure 5.5: AbExp finds more significant gene-trait associations than LOFTEE. (a) Number of genes associated with different traits discovered using AbExp in all tissues compared to using LOFTEE. **(b)** Total number of significant gene-trait associations discovered by different models.

5.3 Phenotype prediction

After demonstrating the enhanced sensitivity for gene-discovery of RVAT with AbExp, I proceeded to evaluate its usefulness in phenotype prediction. For this purpose, I used the remaining third of the dataset to construct gradient boosted decision trees (section 2.7.4) to predict traits using AbExp scores and the number of LOFTEE variants identified in the first two-thirds of the dataset.

5.3.1 Model training and evaluation

For a common variant-based model, I included as predictor variables sex, age, age^2, age times sex, age^2 times sex, the 20 first genetic principal components, and a polygenic risk score predicting the trait (see also section 5.2). I did not include lead trait-associated variants as these would increase the number of predictor variables excessively. For the AbExp-based models, I additionally included the lowest AbExp-predicted z-score across all rare variants for each of the 48 tissues, for all trait-associated genes discovered during RVAT. For the LOFTEE-based models, I used the number of rare LOFTEE-positive variants, for all trait-associated genes discovered during RVAT, in addition to the predictor variables of the model based on common variants only.

For training, I split the remaining third of the dataset across individuals into five cross-validation folds, ensuring an approximately equal distribution of measurements (fig. 5.2). To train and evaluate a prediction model, the model was trained six times on five of these folds and evaluated on the held-out fold, each time using a different held-out fold as the validation set. For all prediction models, I used gradient-boosted trees[59] from the LightGBM[74] framework with default parameters (section 2.7.4). Here, Jonas Lindner helped with the implementation, investigation and visualization of the phenotype prediction models.

5.3.2 AbExp affects the prediction of extreme phenotypes

The predictions of the AbExp-based models rarely deviated from predictions of the common variant-based models. For example, in Alanine aminotransferase less than 0.3% of all individuals had predictions differing by more than 1 standard deviation of the population trait distribution (fig. 5.6a). Individuals with deviating predictions tended to have trait values that deviated significantly from the population average, as can be seen for Alanine Aminotransferase in fig. 5.6b. This suggests that the model incorporating AbExp scores is particularly advantageous in predicting extreme phenotypes that cannot be adequately explained by common variants alone.

In the evaluation on held-out data, the phenotype prediction model leveraging AbExp scores significantly enhanced the explained variation (R^2) in 50% of the traits compared to the model relying on LOFTEE. There were no traits for which the AbExp-based model would significantly reduce the observed R^2 compared to the LOFTEE-based model (fig. 5.7a). Moreover, when considering individuals deviating by more than one standard

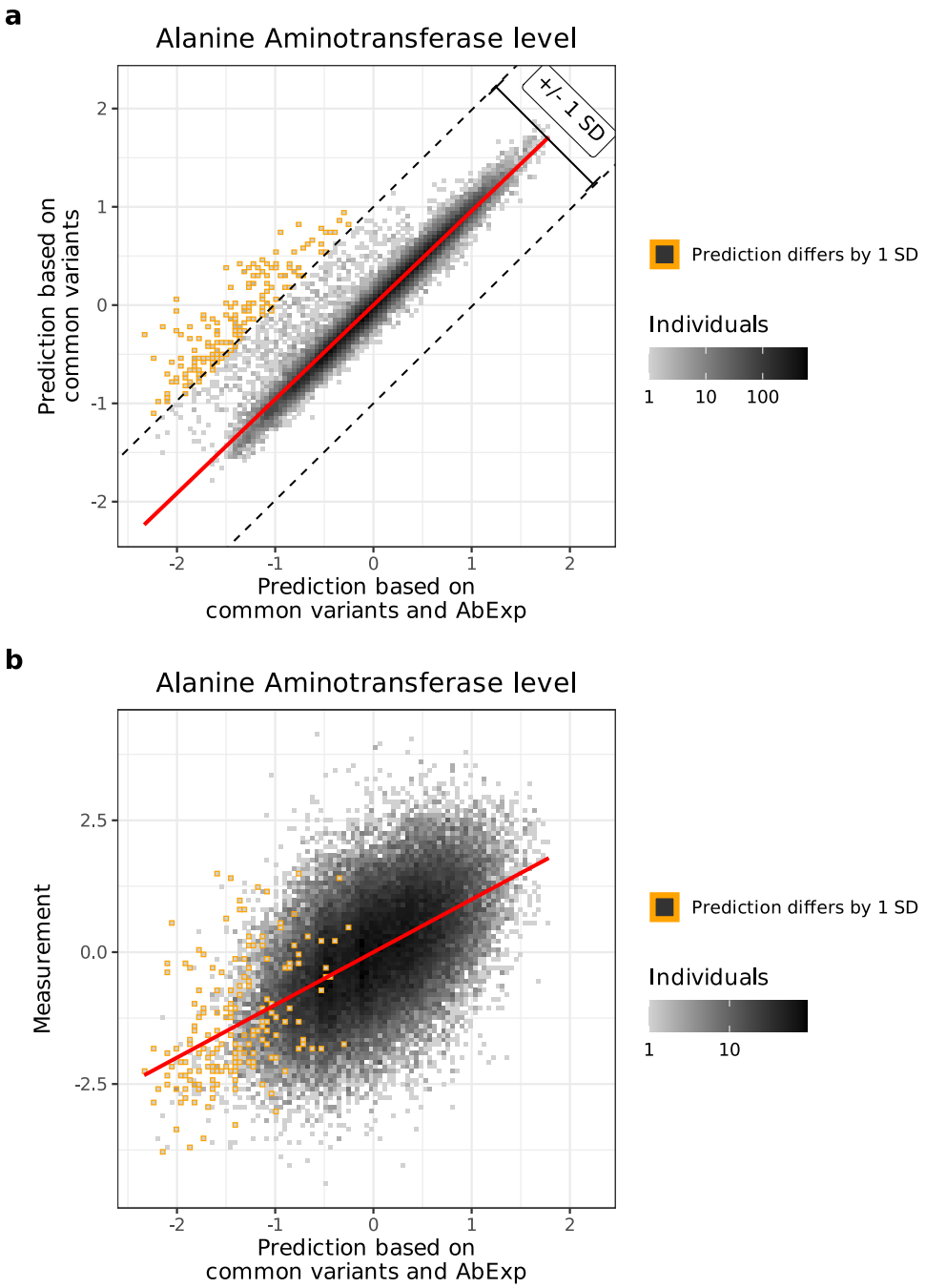

Figure 5.6: Inclusion of AbExp scores affects the prediction of low Alanine aminotransferase levels. **(a)** Alanine aminotransferase level predicted using a model solely based on common variants (y-axis) against predictions using a model based on common variants and AbExp scores (x-axis). Individuals whose predictions differed by more than one standard deviation of the population trait distribution are marked in orange. **(b)** Alanine aminotransferase measurements against predictions based on common variants and AbExp scores. Orange data points as in (a).

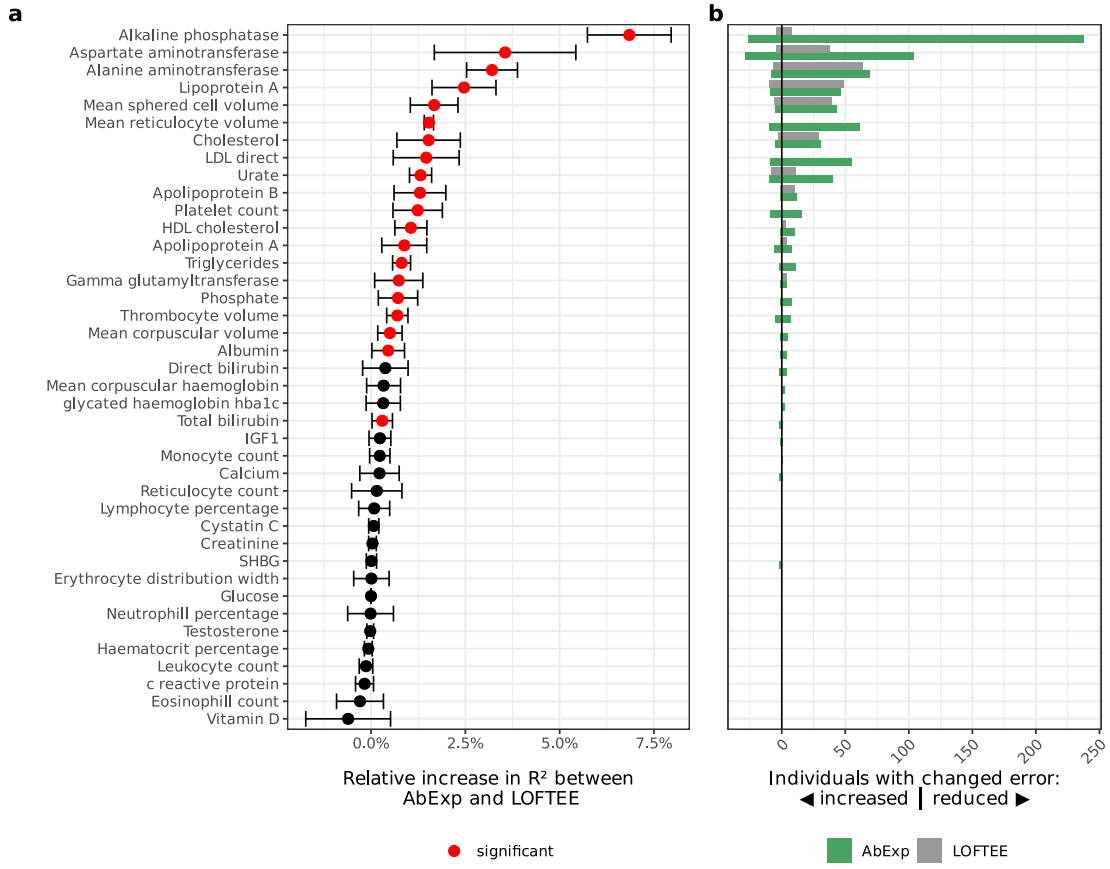

Figure 5.7: AbExp improves phenotype prediction over LOFTEE. (a) Relative R^2 increase between AbExp-based and LOFTEE-based predictions across traits. Traits with a significant difference between both models are marked red (two-sided paired t-test, nominal $P < 0.05$). Error bars show the standard deviation among 5 cross-validation folds. **(b)** Positive bars show the number of individuals with an error reduced by at least one standard deviation in the trait scale and therefore improved prediction, negative bars show the number of individuals with an error increased by at least one standard deviation in the trait scale and therefore worse prediction of the AbExp-based model (green) and the LOFTEE-based model (grey).

deviation from the common variant-based model, the AbExp-based model exhibited improved prediction for 784 individuals across 40 blood traits, while the LOFTEE-based model only enhanced prediction for 259 individuals (fig. 5.7b).

Notably, the benefits of using AbExp scores were consistently evident even when elastic net regularised linear regression (section 2.7.3) was used to predict phenotypes rather than gradient-boosted trees, although this resulted in slightly less accurate predictions of phenotypes (fig. S2). Here, elastic net regression models were trained using the ElasticNetCV implementation of the Scikit-Learn framework[15], which uses grid search to automatically determine the optimal hyperparameters.

5.4 Summary

In this chapter, I demonstrated how AbExp and LOFTEE can be used in rare variant gene association testing and phenotype prediction on 40 blood traits in the UK Biobank. AbExp-based methods detected more significant gene-trait associations and improved the phenotype predictions of more individuals than LOFTEE-based methods, especially in individuals with extreme phenotypes.

6 Discussion

6.1 Benchmarking aberrant gene expression prediction in human tissues

In the beginning of my thesis I have established the first benchmark for aberrant gene expression across 48 human tissues and tested the predictive value of existing variant annotation tools. I found that the variant annotation tools LOFTEE, AbSplice and CADD showed an enrichment in underexpression outliers. However, the predictive value of these models was limited, with models reaching between 0.5% and 1.1% average precision in median across tissue types.

In the future, this benchmark will provide a solid foundation for the systematic comparison of variant annotation tools on their ability to predict underexpression outliers. In particular, it would be interesting to benchmark methods explicitly developed for predicting gene expression from DNA sequence, such as Enformer[6] or Borzoi[97]. Enformer was trained on 5313 different ENCODE and FANTOM5 tracks from human and mouse, including ChIP-seq for hundreds of transcription factors, DNase-seq which measures genome accessibility; and cap-analysis of gene expression (CAGE) measurements. Borzoi paired the training data from Enformer with human and mouse RNA-seq measurements from the ENCODE project, as well as 2-3 replicates for each GTEx tissue. While sequence-based models of cis-regulation are typically trained on the entire spectrum of expression levels, this work focused on extreme expression variations. Models trained to predict gene expression globally might not fully capture these extremes. Furthermore, the biological mechanisms underlying extreme expression variations may differ from those leading to moderate ones.

This benchmark has limitations. As shown in fig. 3.6a, 41% of underexpression outliers and 57% of overexpression outliers do not have rare variants within ± 5,000 bp of the gene. While these outliers might be caused by variants outside of this region, they could as well be caused by effects that a sequence-based model cannot explain, such as environmental or epigenomic factors, or sequencing artifacts. It is also possible that there are artifacts caused by the outlier caller. Further, I found that 90% of the overexpression outliers and 78% of the underexpression outliers are singletons (fig. 3.4). With only 40-44% of these singletons having a rare variant within 5,000 bp of the gene (fig. 3.6b), it is unclear which of the singletons are caused by calling artifacts or other non-sequence effects. Outlier calling might be improved by joint calling of all tissues, with the tissue as a covariate, or by including environmental factors as covariates. Also, correlation with

epigenomic measurements could allow to remove (non-)outliers from the benchmark that might be caused by abnormal epigenetic states. Finally, this study primarily focuses on underexpression outliers and excludes genes with a low expected expression rate. To effectively benchmark overexpression outliers, it would be useful to keep such genes since variants might activate gene expression.

6.2 AbExp: Predicting aberrant gene underexpression across human tissues

Building on the benchmark results, I developed AbExp, a method to predict aberrantly underexpressed genes in human tissues from DNA sequence by integrating existing variant annotations with tissue-specific gene expression variability and transcript isoform composition. AbExp outperformed existing variant annotation tools between 6-fold and 18-fold in median average precision across tissues on this task, and this performance improvement was consistent across independent datasets. Combining AbExp scores with gene expression measurements from clinically accessible tissues to predict aberrant expression in other tissues yielded a further performance increase by 2-fold over AbExp.

AbExp predicts a continuous, tissue-specific z-score of gene expression. On average, AbExp predicted 1.2 high-confidence and 5.7 low-confidence underexpressed genes per individual. Even with a low confidence threshold, AbExp was able to distinguish pathogenic variants from benign and likely benign variants in the ClinVar database with a high precision of 98.1%. This highlights the potential of AbExp as a tool for the prioritization of pathogenic variants.

The development of AbExp also revealed interesting insights into the underlying biology of gene expression outliers. While confirming the relevance of nonsense-mediated mRNA decay among underexpressed genes[25], this work also highlighted the importance of taking tissue-specific transcript isoforms into account. Furthermore, the impact of a certain type of genetic variant correlates with the expression variability of the gene it affects. On one hand, the same change in gene expression is more likely to lead to an expression outlier in a tightly controlled gene with low expression variability among the population. On the other hand, the change in gene expression caused by a variant is not uniform across genes. Instead, the same type of variant tends to change the expression of genes with a low variability stronger than in other genes, suggesting that regulatory buffering mechanisms partially recover gene expression in tightly controlled genes.

6.2.1 AbExp assumes that outliers are caused by rare variants within gene regions

AbExp relies on a number of assumptions that do not necessarily hold. First, it assumes that an underexpression outlier is caused by a rare variant. However, it is also possible that an expression outlier is caused by a rare combination of two or more frequently

occurring variants. Conversely, damage caused by one variant might be recovered by another variant, e.g. two frameshift variants where one variant can restore the correct reading frame after the other variant causes a disruption. Incorporating combinations of variants would require a more complex model that also considers phasing of the variants.

Also, AbExp does not consider whether a variant is heterozygous or homozygous. In theory, a variant should have twice the effect if it affects both copies of a gene. An improved model might use this variant zygosity as additional information on outlier prediction, assuming there is a sufficient number of homozygous variants in the dataset. However, while individuals with a heterozygous high-impact variant might be viable due to a second unaffected copy of the gene, homozygous high-impact variants might be depleted in the population when both copies of the gene are affected, limiting the data available for training.

Further, AbExp only considers variants within ± 5,000 bp of protein-coding genes. This limited window misses longer-range interactions such as enhancers and silencers and therefore variants possibly causing expression outliers. Predicting the gene expression impact of enhancer and silencer variants requires evaluating whether the variant impairs the enhancer or silencer and further the tissue-specific regulatory impact of the enhancer on the gene of interest. While current sequence-based methods are in general unable to accurately predict the gene expression effect of enhancer variants, combining these with tissue-specific maps of enhancer-gene interactions might provide valuable information for expression outlier prediction[72].

Finally, AbExp does not consider the effects of trans-acting gene regulation. For example, aberrant expression of a specific transcription factor might affect gene expression of a large number of other genes. Also, the cause of the observed expression buffering in genes with a low expression variability is unclear. Modeling such effects would require a very different approach that captures regulatory networks. In this context, it would also be sensible to expand aberrant expression prediction to non-coding genes. Non-coding RNA can control gene expression by attaching to coding RNA, along with certain proteins, to break down the coding RNA. Non-coding RNA may also recruit proteins to modify histones, changing the accessibility of other genes and therefore their expression rate[68].

6.2.2 Additional annotations could improve AbExp

AbExp could be improved in multiple ways. As of now, the AbExp pipeline relies on the CADD plugin within Ensembl VEP, which annotates only pre-computed CADD scores, potentially overlooking annotations for rare variants. To ensure comprehensive annotation of all rare variants, it would be necessary to directly predict CADD scores using the original CADD pipeline.

Although AbExp considers structural variants, it only incorporates transcript ablations caused by long deletions. However, as outlined in section 1.3.3, there are other types of structural variants such as copy number variations that are strongly enriched

among expression outliers. More types of structural variants could be interpreted with specialized tools such as CADD-SV, a tool that scores the deleteriousness of large (>50 bp) deletions, insertions, and duplications[78]. Also, one could investigate the copy number frequency among the global population using the latest GnomAD v4 release[70, 73]. Another type of variant to consider are short tandem repeat (STR) variations, i.e. variations in the number of repeat units within STR regions of the genome. STR variations have been associated with the expression of nearby genes in GTEx[44] and are a known cause of rare diseases[37, 117].

Further improvements could be made in the interpretation of promoter variants. In GTEx, I found 188 rare variants in a -100 to $+50$ bp window around the transcription start sites in underexpressed genes, with an enrichment of up to 15% among rare promoter variants (fig. 6.1). AbExp interprets the impact of these variants based on the CADD score only, which is neither tissue-specific nor specialized to predict the impact on gene expression. A recent study has shown that Enformer, a deep learning model that predicts gene expression from DNA sequences by integrating long-range interactions, accurately captures gene expression determinants in promoters on GTEx, suggesting that the inclusion of Enformer predictions might improve the overall predictive value of AbExp[72].

In GTEx, there is also a notable enrichment of rare variants within a certain distance to transcription termination sites (fig. 6.2). In total, I identified 73 rare variants within a 60 bp window upstream of the transcription termination sites of underexpressed genes, with an enrichment of up to 27% among rare polyadenylation region variants. This enrichment could be attributed to a major determinant of the polyadenylation cut site, namely the presence of canonical sequence elements located approximately 50 base pairs upstream of the cut site[13]. Variants destroying those sequence elements might impair the successful transcription of these genes. One could explore whether these variants predict underexpression, for instance, by incorporating a binary variable into AbExp indicating the presence of a rare variant within a 60bp window upstream of the TTS. Alternatively, sequence-based methods like APARENT and APARENT2[13, 96], which quantitatively predict alternative polyadenylation, could be integrated as additional annotations in AbExp.

Although missense variants mainly affect the functionality of the protein, changes in the encoded protein sequence can also lead to ribosomal stalling and mRNA degradation by the ribosome-mediated quality control and therefore a reduction in gene expression levels (fig. 6.3)[125]. We suspect that AlphaMissense[19], a recent model for predicting the impact of missense variants, could be predictive of such effects. AlphaMissense scores could be integrated into AbExp to test this hypothesis. Also, one could investigate the predictive value of mRNA secondary structure prediction or codon optimality-mediated decay[8, 51].

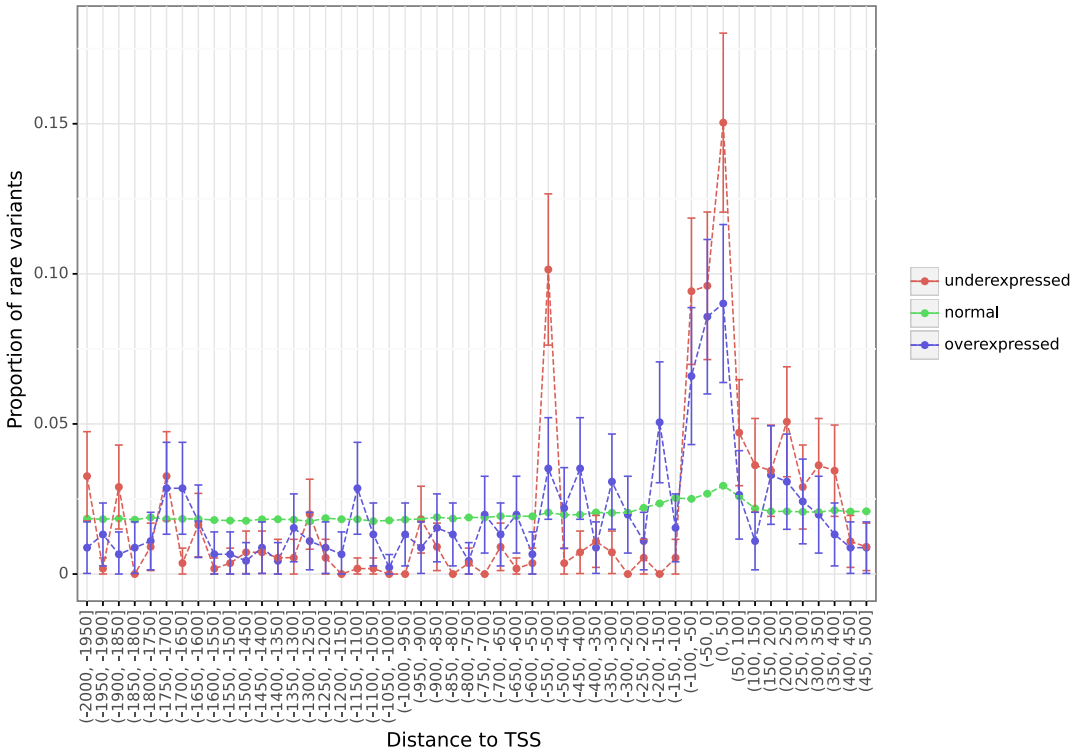

Figure 6.1: Enrichment of rare promoter variants relative to the transcription start site (TSS) in the GTEx benchmark dataset. The y-axis shows the proportion of rare variants in a certain distance interval to the TSS (x-axis) among all rare variants within 2,000 upstream to 500 bp downstream of a TSS, grouped by underexpression outliers, overexpression outliers, and non-outliers. Error bars mark 95% binomial confidence intervals.

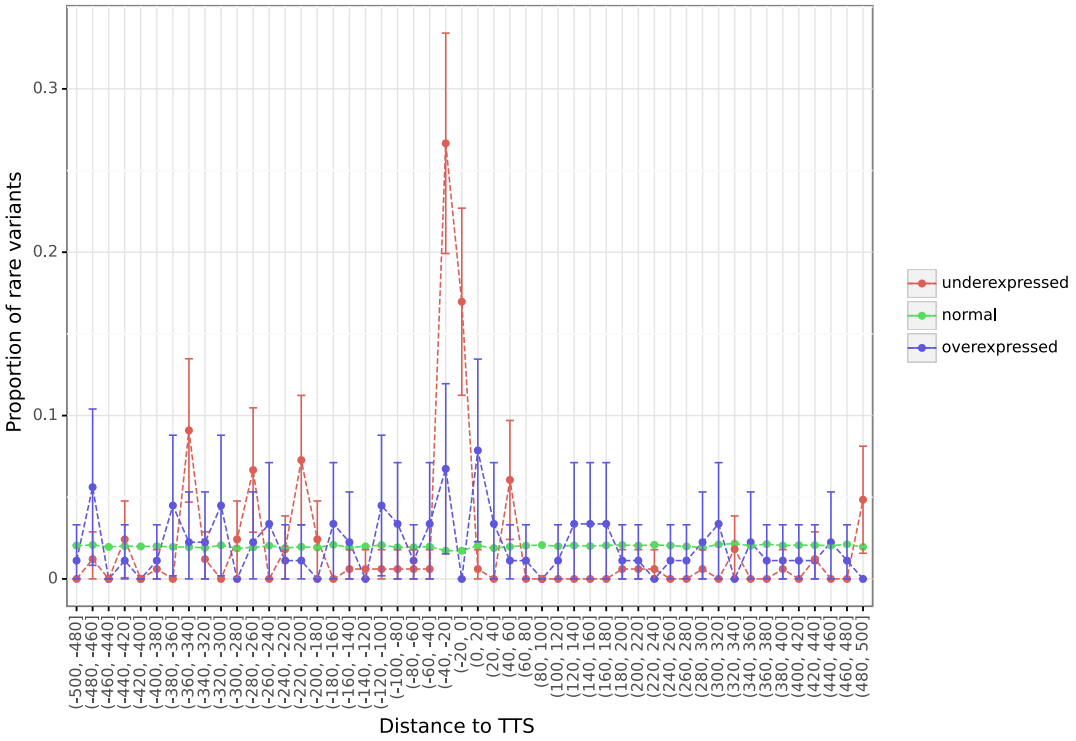

Figure 6.2: Enrichment of rare promoter variants relative to the transcription termination site (TTS) in the GTEx benchmark dataset. The y-axis shows the proportion of rare variants in a certain distance interval to the TTS (x-axis) among all rare variants within ± 500 bp of a TTS, grouped by underexpression outliers, overexpression outliers, and non-outliers. Error bars mark 95% binomial confidence intervals.

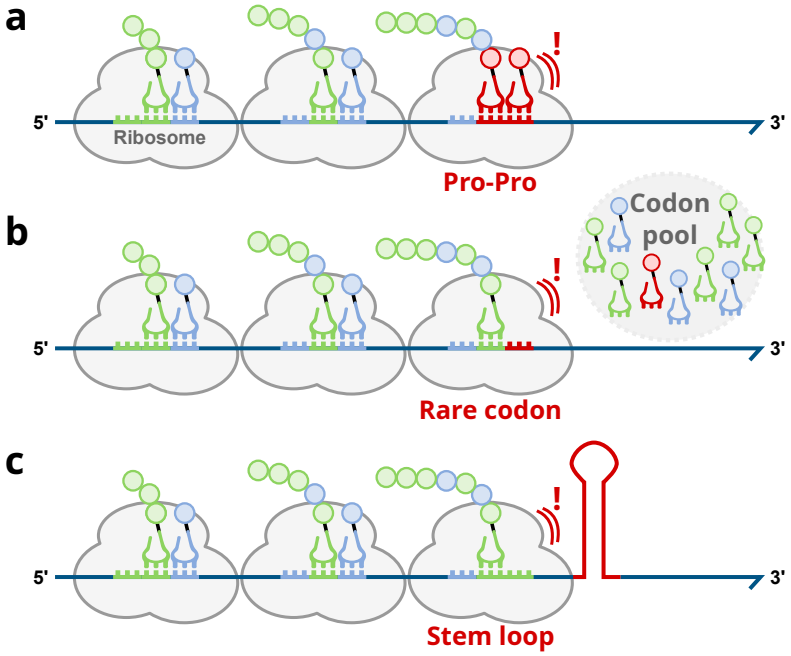

Figure 6.3: Potential causes of ribosome stalling. Ribosome stalling can be caused by **(a)** specific amino acid combinations, e.g. Prolin-Prolin pairs, **(b)** bad codon optimality, i.e. the usage of rare codons, or **(c)** stem-loops or pseudoknots in the mRNA structure.

6.2.3 Aberrant expression prediction in more settings

AbExp predicts aberrant gene expression in 48 human tissues. Extending its predictive ability to other tissues is comparatively simple as it does not require retraining of the model. Instead, it would be theoretically sufficient to compute isoform composition, average gene expression, coefficients of gene expression variation, and SpliceMaps based on a representative set of RNA-seq samples. One could explore whether such custom annotations influence performance, and how many tissue samples are required to obtain a good reference for predicting aberrant expression in tissues that are not covered by GTEx. For example, predictions in ALS might be improved by creating specialized annotations for induced pluripotent stem cells, instead of using nerve tissue from GTEx as a reference.

Further, the tissues measured in GTEx are mixtures of various cell types and therefore also the measured gene expression. For example, lung tissue is composed of various cell types such as alveolar cells, bronchial epithelium, alveolar macrophages, endothelial cells, and interstitial cells[71]. Developing a version of AbExp based on single-cell transcriptomics measurements could refine its predictions to cell-type resolution.

Although expression patterns vary widely in different tissues and individuals, they remain relatively stable in adulthood. However, gene expression patterns change during development and play a central role in health and disease milestones, such as the onset

or end of growth spurts, the onset of puberty, and the onset of child-specific conditions and diseases that manifest later in life. The developmental Genotype-Tissue Expression (dGTEx) project, an emerging resource of gene expression measurements among kids in different developmental stages[27], could provide a valuable resource to study aberrant expression patterns during human development.

AbExp could also be applied to somatic mutations in cancer tissues and cell lines, potentially aiding in cancer type classification and driver gene discovery. Conversely, cancer cells represent natural perturbation experiments and, thus, gene expression measurements in cancer cell lines might be useful to further test and improve AbExp.

6.3 The UK Biobank rare variant association testing and phenotype prediction study can be improved

Finally, in the third part of my thesis, I demonstrated the application of AbExp in rare variant association testing and phenotype prediction on 40 blood traits of the UK Biobank. AbExp scores offer supplementary information beyond the state-of-the-art putative loss of function classifier LOFTEE, significantly enhancing rare variant gene association testing as well as phenotype prediction.

While confirming that rare expression outlier associated variants are predictive of phenotypic traits, this study goes beyond the state-of-the-art of Smail et al.[132] that is restricted to rare variants found in GTEx. Since AbExp generalizes to unseen variants, AbExp allowed to evaluate the gene expression impact of any variant within 5,000 bp of genes in the UK Biobank. This is especially relevant concerning rare variant-based studies, given that many rare variants, especially those unique to a single individual, are unlikely to be found in GTEx. Expanding the analysis to include variants beyond those found in GTEx increases the likelihood of detecting significant associations and enhances the potential to accurately predict phenotypes.

While the UK Biobank case study was performed on a set of 200,000 whole-exome sequencing samples, recent additions to the UK Biobank include whole-genome sequencing data for all 500,000 participants in the UK Biobank. Leveraging this additional data would improve both gene discovery and phenotype prediction, not only due to the increased sample size, but also due to the improved prediction of aberrant underexpression as whole genome sequencing covers, in contrast to whole exome sequencing, also other relevant regions such as deep intronic or promoter regions. The case study could also be extended to include other traits such as diseases. Further, better modeling approaches that also consider variant combinations might improve the detection of gene-trait associations and phenotype prediction[22]. Finally, it would be interesting to study which tissues showed the most significant gene-trait associations.

6.4 Conclusion and outlook

In summary, the development of a DNA sequence-based method for predicting aberrant gene expression in multiple tissues, which can also generalize to unseen variants, represents a significant advancement in the ability to identify and understand the genetic underpinnings of human traits and diseases. Hopefully, the benchmark and algorithms presented in this study will encourage further research in this area and assist in developing and validating methods for predicting the impact of large-effect variants on the human transcriptome.

Despite being a critical determinant of protein abundance, gene expression is not the sole factor. To fully understand the functional impact of genetic variants on proteins, it is necessary to consider also post-transcriptional effects, such as the availability of resources for protein biosynthesis[98]. With the UK Biobank Pharma Proteomics Project (UKB-PPP), a study of plasma proteomic profiles from 54,306 UK Biobank participants, arises the unique possibility to investigate these effects. Similar to this study on aberrant gene expression on GTEx, a comparable approach could be applied to the UKB-PPP dataset using PROTRIDER, a specialized tool for detecting aberrant protein expression similar to OUTRIDER[80]. Based on this, one could establish a benchmark to test and develop tools predicting the impact of genetic variants at the protein level.

Besides proteins, cells also harbor other disease-relevant metabolites such as sugars, fatty acids, lipids, and steroids. For example, the deregulation of metabolic pathways plays an important role in oncogenesis[57]. Here, exploring aberrant metabolomic profiles, e.g. predicted from aberrant gene expression, could provide valuable insights[17].

A Data and code availability

The aberrant expression benchmark dataset, isoform proportions, and the expected gene expression in GTEx v8 are available as open-access in the Zenodo repository[62] with DOI: 10.5281/zenodo.8427312.
A Snakemake pipeline to calculate AbExp predictions can be found at:
https://github.com/gagneurlab/abexp
See also section 4.9. The source code for the UK Biobank rare-variant association study and phenotype prediction can be found in the following repositories:

- Main analysis pipeline: https://github.com/gagneurlab/abexp-ukbb-trait-analysis
- Variant clumping: https://github.com/gagneurlab/abexp-ukbb-variant-clumping
- Polygenic risk score calculation: https://github.com/gagneurlab/abexp-ukbb-prs

B Supplementary figures

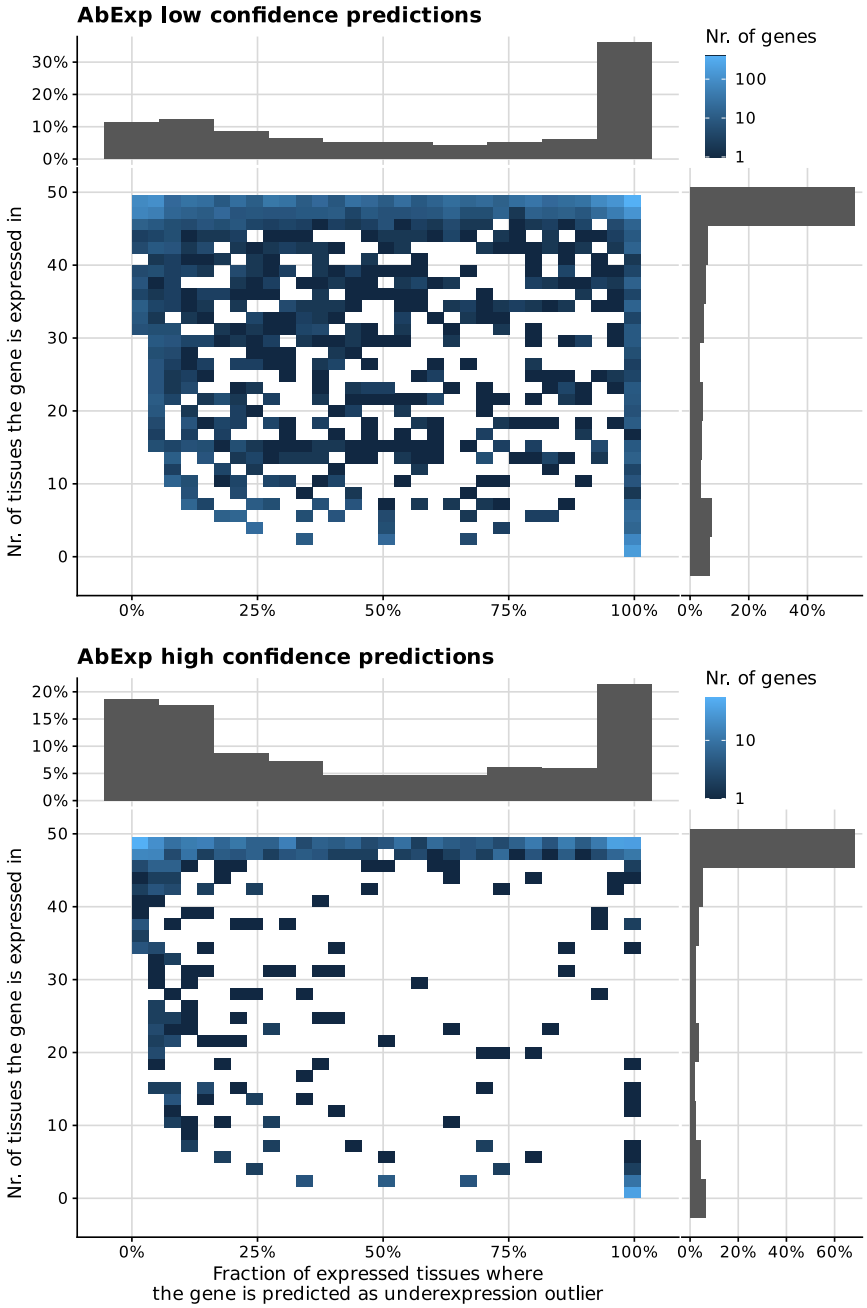

Figure S1: Relationship between the fraction of predicted multi-tissue outliers and the number of expressed tissues among AbExp low and high confidence predictions. The color denotes the number of genes AbExp predicted with low or high confidence as underexpressed in at least one tissue in the GTEx dataset. The x-axis shows the fraction of tissues where the gene is predicted as underexpression outlier. The y-axis shows the number of tissues the gene is expressed in. About 9-10% of the predicted genes are expressed in less than five tissues, contributing 7-8% of genes predicted as underexpression outliers in at least 90% of tissues. Notably, many high-confidence predictions are outliers in only few tissues.

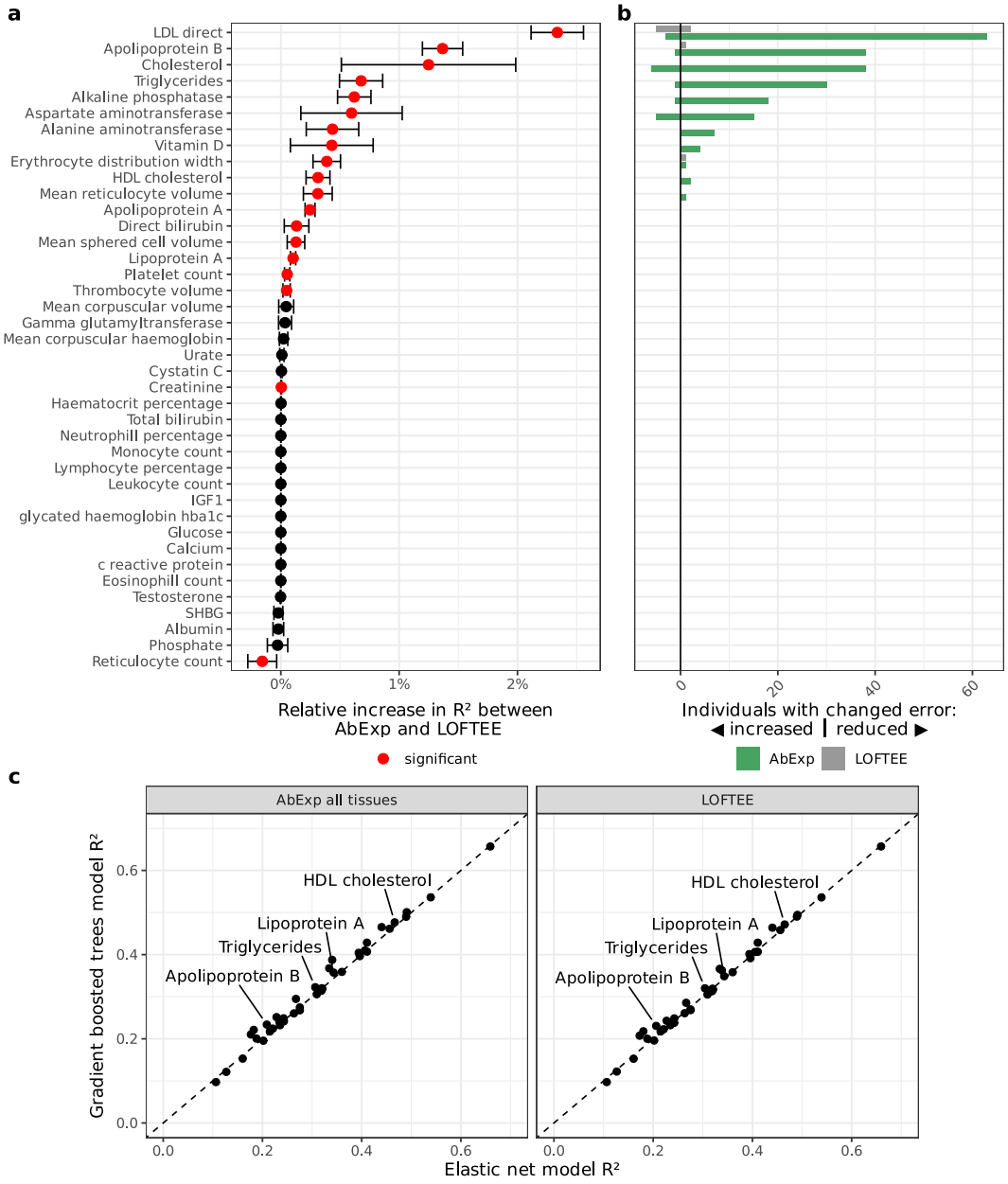

Figure S2: AbExp improvements over LOFTEE in phenotype prediction also holds with elastic net regularised linear regression predictions. **(a)** Relative R^2 increase between AbExp-based and LOFTEE-based predictions across traits. Traits with a significant difference between both models are marked red (two-sided paired t-test, nominal $P < 0.05$). Error bars show the standard deviation among 5 cross-validation folds. **(b)** Positive bars show the number of individuals with an error reduced by at least one standard deviation in the trait scale and therefore improved prediction, negative bars show the number of individuals with an error increased by at least one standard deviation in the trait scale and therefore worse prediction of the AbExp-based model (green) and the LOFTEE-based model (grey). **(c)** R^2 of gradient boosted trees models against R^2 of elastic net models across traits, when using AbExp scores or LOFTEE. All data presented in this figure are computed on held-out folds of a 5-fold cross-validation within a third of the UKBB data.

List of Figures

1.1 Warfarin mechanism of action . . . 2
1.2 Distribution of reported cases among rare diseases in the Orphanet knowledge base . . . 3
1.3 GWAS links single genetic variants to heritable traits . . . 5
1.4 Variant effect size against variant allele frequency . . . 6
1.5 Distribution of rare variants among 635 individuals . . . 7
1.6 An enhancer-promoter interaction within a topologically associating domain formed by CTCF and cohesin . . . 9
1.7 Proportion of gene expression outliers potentially explained by different classes of rare variants in GTEx v6p . . . 12

2.1 Overview of types of genetic variation . . . 16
2.2 The central dogma of biology . . . 17
2.3 The standard RNA codon table organized in a wheel . . . 18
2.4 Overview of nonsense-mediated decay pathway . . . 19
2.5 Premature termination codons (PTCs) in certain regions of the gene do not induce nonsense-mediated decay (NMD) . . . 20
2.6 Next-generation sequencing . . . 21
2.7 Read sequence alignment to the reference genome . . . 22
2.8 Context-dependent outlier detection with OUTRIDER . . . 24
2.9 Illustration of a linear regression model with one predictor variable . . . 27
2.10 Contour plot of least squares cost function and various regularization methods for a two-dimensional model . . . 29
2.11 Building an ensemble model of decision trees with gradient boosting . . . 31
2.12 Example of two alternative models fitting some observed values . . . 34
2.13 Variant consequences calculated by Ensembl VEP . . . 35
2.14 Tissue sampling sites used in the GTEx project . . . 37
2.15 Sample size across major ExAC/gnomAD releases . . . 40

3.1 A benchmark for aberrant gene expression prediction across human tissues 42
3.2 Only 0.9% of GTEx transcriptome sequencing samples have more than 50 outliers . . . 44
3.3 Most RNA-seq samples in the GTEx dataset are from skin or brain tissues 47
3.4 90% of the overexpression outliers and 78% of the underexpression outliers are singletons . . . 48

3.5 In most tissues, samples have in median two underexpression and two overexpression outliers . . . 49
3.6 Not all outliers can be explained by variants within gene regions . . . 50
3.7 Enrichment of various variant annotations in expression outliers across tissues . . . 52
3.8 Study design and main findings of AbSplice . . . 54
3.9 AbSplice outperforms state-of-the-art splicing models in predicting aberrant splicing . . . 55
3.10 Performance of various variant annotations in underexpression prediction 57

4.1 Improving aberrant underexpression prediction in human tissues . . . 60
4.2 Cross-validation scheme . . . 60
4.3 Gene expression z-scores are standard-normally distributed . . . 62
4.4 Quantile-mapping of OUTRIDER-fitted fragment count distribution to standard normal distribution . . . 63
4.5 An integrative model outperforms both LOFTEE and CADD on predicting underexpression outliers . . . 64
4.6 Tissue-specific isoform expression in *PSMB10* leads to tissue-specific aberrant gene expression . . . 66
4.7 Most canonical transcript isoforms contribute only a fraction of the total gene expression . . . 67
4.8 Illustration of stop-gain consequence score calculation for two variants . . . 68
4.9 Accounting for tissue-specific isoform expression improves predictions . . . 69
4.10 The outlier state depends on the tissue-specific biological coefficient of variation . . . 71
4.11 Fold-changes of the same class of variants are correlated with expression variability . . . 72
4.12 Incorporating the tissue-specific gene expression variability improves predictions . . . 73
4.13 AbExp combines various variant and tissue annotations to predict aberrant gene expression and outperforms LOFTEE by about sevenfold . . . 74
4.14 Performance of AbExp replicates on independent datasets . . . 78
4.15 AbExp predicts on average per individual 1.2 low-confidence and 5.7 high-confidence genes to be underexpressed in at least one tissue . . . 80
4.16 AbExp predictions are tissue-specific . . . 81
4.17 Fraction of variant types among AbExp high-impact predictions . . . 82
4.18 Gene expression z-score versus AbExp-DNA predictions for various types of variants . . . 83
4.19 AbExp predicts pathogenic variants with high precision . . . 85
4.20 Gene expression correlates between clinically accessible tissues and non-accessible tissues . . . 87

4.21 Combining RNA-seq measurements from clinically accessible tissues with AbExp improves the prediction performance 89

5.1 Rare variant association testing and phenotype prediction with AbExp . . 92
5.2 Illustration of data splitting for rare variant association testing and phenotype prediction 93
5.3 Accounting for common variation in RVAT 94
5.4 P-values of all models are calibrated 97
5.5 AbExp finds more significant gene-trait associations than LOFTEE 98
5.6 Inclusion of AbExp scores affects the prediction of low Alanine aminotransferase levels 100
5.7 AbExp improves phenotype prediction over LOFTEE 101

6.1 Enrichment of rare promoter variants relative to the transcription start site (TSS) in the GTEx benchmark dataset 107
6.2 Enrichment of rare promoter variants relative to the transcription termination site (TTS) in the GTEx benchmark dataset 108
6.3 Potential causes of ribosome stalling 109

S1 Relationship between the fraction of predicted multi-tissue outliers and the number of expressed tissues among AbExp low and high confidence prediction 116
S2 AbExp improvements over LOFTEE in phenotype prediction also holds with elastic net regularised linear regression predictions 117

List of Tables

3.1 GTEx dataset filtering 46

4.1 Mitochondrial disease dataset filtering 76
4.2 ALS dataset filtering 77

5.1 Rare variant association test models 94
5.2 List of blood traits and corresponding PGS catalog IDs of polygenic risk scores 95

Acronyms

AP	average precision.
AUPRC	area under the precision-recall curve.
CNV	copy number variation.
DNA	deoxyribonucleic acid.
DNA-seq	RNA sequencing.
FDR	False Discovery Rate.
INDEL	insertion or deletion.
mRNA	messenger RNA.
NGS	next-generation sequencing.
NMD	nonsense-mediated decay.
pre-mRNA	pre-mature messenger RNA.
PTC	premature termination codon.
RNA	ribonucleic acid.
RNA-seq	RNA sequencing.
SNV	single nucleotide variant.
STR	short tandem repeat.
SV	structural variant.
tRNA	transfer ribonucleic acid.
TSS	transcription start site.

References

[1] Sergey Aganezov et al. "A Complete Reference Genome Improves Analysis of Human Genetic Variation." In: *Science* (2022). DOI: 10.1126/science.abl3533.

[2] Joseph K Aicher et al. "Mapping RNA Splicing Variations in Clinically-Accessible and Non-Accessible Tissues to Facilitate Mendelian Disease Diagnosis Using RNA-seq." In: *Genetics in medicine : official journal of the American College of Medical Genetics* (2020). DOI: 10.1038/s41436-020-0780-y.

[3] Bruce Alberts. *Molecular Biology of the Cell*. 6th ed. Garland Science, 2017. 1465 pp. ISBN: 978-1-317-56375-4.

[4] Shanika L. Amarasinghe et al. "Opportunities and Challenges in Long-Read Sequencing Data Analysis." In: *Genome Biology* (2020). DOI: 10.1186/s13059-020-1935-5.

[5] Adam Auton et al. "A Global Reference for Human Genetic Variation." In: *Nature* (2015). DOI: 10.1038/nature15393.

[6] Žiga Avsec et al. "Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions." In: *Nature Methods* (2021). DOI: 10.1038/s41592-021-01252-x.

[7] Daniel M Bader et al. "Negative Feedback Buffers Effects of Regulatory Variants." In: *Molecular Systems Biology* (2015). DOI: 10.15252/msb.20145844.

[8] Haneui Bae and Jeff Coller. "Codon Optimality-Mediated mRNA Degradation (COMD): Linking Translational Elongation to mRNA Stability." In: *Molecular cell* (2022). DOI: 10.1016/j.molcel.2022.03.032.

[9] Francisco E Baralle and Jimena Giudice. "Alternative Splicing as a Regulator of Development and Tissue Identity." In: *Nature reviews. Molecular cell biology* (2017). DOI: 10.1038/nrm.2017.27.

[10] Emily G. Baxi et al. "Answer ALS, a Large-Scale Resource for Sporadic and Familial ALS Combining Clinical and Multi-Omics Data from Induced Pluripotent Cell Lines." In: *Nature Neuroscience* (2022). DOI: 10.1038/s41593-021-01006-0.

[11] Yoav Benjamini and Daniel Yekutieli. "The Control of the False Discovery Rate in Multiple Testing under Dependency." In: *The Annals of Statistics* (2001). DOI: 10.1214/aos/1013699998.

[12] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer, 2006. 738 pp. ISBN: 978-0-387-31073-2.

[13] Nicholas Bogard et al. "A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation." In: *Cell* (2019). DOI: 10.1016/j.cell.2019.04.046.

[14] Felix Brechtmann et al. "OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data." In: *The American Journal of Human Genetics* (2018). DOI: 10.1016/j.ajhg.2018.10.025.

[15] Lars Buitinck et al. "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project." In: *arxiv* (2013). DOI: 10.48550/arXiv.1309.0238. preprint.

[16] William S. Bush and Jason H. Moore. "Chapter 11: Genome-Wide Association Studies." In: *PLOS Computational Biology* (2012). DOI: 10.1371/journal.pcbi.1002822.

[17] Maria Vittoria Cavicchioli et al. "Prediction of Metabolic Profiles from Transcriptomics Data in Human Cancer Cell Lines." In: *International Journal of Molecular Sciences* (2022). DOI: 10.3390/ijms23073867.

[18] Christopher C Chang et al. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." In: *GigaScience* (2015). DOI: 10.1186/s13742-015-0047-8.

[19] Jun Cheng et al. "Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense." In: *Science* (2023). DOI: 10.1126/science.adg7492.

[20] Jun Cheng et al. "MMSplice: Modular Modeling Improves the Predictions of Genetic Variant Effects on Splicing." In: *Genome Biology* (2019). DOI: 10.1186/s13059-019-1653-z.

[21] Yongwook Choi et al. "Comparison of Phasing Strategies for Whole Human Genomes." In: *PLOS Genetics* (2018). DOI: 10.1371/journal.pgen.1007308.

[22] Brian Clarke et al. "Integration of Variant Annotations Using Deep Set Networks Boosts Rare Variant Association Genetics." In: *bioRxiv* (2023). DOI: 10.1101/2023.07.12.548506. preprint.

[23] GTEx Consortium. "The Genotype-Tissue Expression (GTEx) Project." In: *Nature genetics* (2013). DOI: 10.1038/ng.2653.

[24] Beryl B. Cummings et al. "Improving Genetic Diagnosis in Mendelian Disease with Transcriptome Sequencing." In: *Science Translational Medicine* (2017). DOI: 10.1126/scitranslmed.aal5209.

[25] Beryl B. Cummings et al. "Transcript Expression-Aware Annotation Improves Rare Variant Interpretation." In: *Nature* (2020). DOI: 10.1038/s41586-020-2329-2.

[26] Ruebena Dawes, Himanshu Joshi, and Sandra T. Cooper. "Empirical Prediction of Variant-Activated Cryptic Splice Donors Using Population-Based RNA-Seq Data." In: *Nature Communications* (2022). DOI: 10.1038/s41467-022-29271-y.

[27] *dGTEx*. URL: https://dgtex.org/ (visited on 01/30/2024).

[28] Dongsheng Duan et al. "Duchenne Muscular Dystrophy." In: *Nature Reviews Disease Primers* (2021). DOI: 10.1038/s41572-021-00248-3.

[29] Sarah L. Dugan et al. "New Recessive Truncating Mutation in LTBP3 in a Family with Oligodontia, Short Stature, and Mitral Valve Prolapse." In: *American Journal of Medical Genetics Part A* (2015). DOI: 10.1002/ajmg.a.37049.

[30] Karen Eilbeck et al. "The Sequence Ontology: A Tool for the Unification of Genome Annotations." In: *Genome Biology* (2005). DOI: 10.1186/gb-2005-6-5-r44.

[31] Hjörleifur Einarsson et al. "Promoter Sequence and Architecture Determine Expression Variability and Confer Robustness to Genetic Variants." In: *eLife* (2022). DOI: 10.7554/eLife.80943.

[32] *Ensembl VEP Plugins: NMD*. GitHub. Apr. 4, 2023. URL: https://github.com/Ensembl/VEP_plugins/blob/release/111/NMD.pm (visited on 02/09/2024).

[33] *Ensembl/Public-Plugins V108*. Version v108. Ensembl Project, Jan. 11, 2024. URL: https://github.com/Ensembl/public-plugins/ (visited on 02/09/2024).

[34] European Parliament and Council. *Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on Orphan Medicinal Products*. 1999.

[35] Warren J. Ewens and Gregory R. Grant. *Statistical Methods in Bioinformatics: An Introduction*. 2nd ed. 2005. Corr. 2nd printing 2005 Edition. New York, N.Y: Springer, 2004. 618 pp. ISBN: 978-0-387-40082-2.

[36] Benjamin Jung Fair et al. "Gene Expression Variability in Human and Chimpanzee Populations Share Common Determinants." In: *eLife* (2020). DOI: 10.7554/eLife.59929.

[37] Sarah Fazal et al. "Repeat Expansions Nested within Tandem CNVs: A Unique Structural Change in GLS Exemplifies the Diagnostic Challenges of Non-Coding Pathogenic Variation." In: *Human Molecular Genetics* (2023). DOI: 10.1093/hmg/ddac173.

[38] Lewis J. Fermaglich and Kathleen L. Miller. "A Comprehensive Study of the Rare Diseases and Conditions Targeted by Orphan Drug Designations and Approvals over the Forty Years of the Orphan Drug Act." In: *Orphanet Journal of Rare Diseases* (2023). DOI: 10.1186/s13023-023-02790-7.

[39] Nicole M. Ferraro et al. "Transcriptomic Signatures across Human Tissues Identify Functional Rare Genetic Variation." In: *Science* (2020). DOI: 10.1126/science.aaz5900.

[40] Farzaneh Fesahat, Fateme Montazeri, and Seyed Mehdi Hoseini. "Preimplantation Genetic Testing in Assisted Reproduction Technology." In: *Journal of Gynecology Obstetrics and Human Reproduction* (2020). DOI: 10.1016/j.jogoh.2020.101723.

[41] Chris Finan et al. "The Druggable Genome and Support for Target Identification and Validation in Drug Development." In: *Science Translational Medicine* (2017). DOI: 10.1126/scitranslmed.aag1166.

[42] Rebecca C. Fitzgerald et al. "The Future of Early Cancer Detection." In: *Nature Medicine* (2022). DOI: 10.1038/s41591-022-01746-x.

[43] Petko P. Fiziev et al. "Rare Penetrant Mutations Confer Severe Risk of Common Diseases." In: *Science* (2023). DOI: 10.1126/science.abo1131.

[44] Stephanie Feupe Fotsing et al. "The Impact of Short Tandem Repeat Variation on Gene Expression." In: *Nature genetics* (2019). DOI: 10.1038/s41588-019-0521-9.

[45] Adam Frankish et al. "GENCODE Reference Annotation for the Human and Mouse Genomes." In: *Nucleic Acids Research* (2019). DOI: 10.1093/nar/gky955.

[46] Jonathan Frazer et al. "Disease Variant Prediction with Deep Generative Models of Evolutionary Data." In: *Nature* (2021). DOI: 10.1038/s41586-021-04043-8.

[47] Laure Frésard et al. "Identification of Rare-Disease Genes Using Blood Transcriptome Sequencing and Large Control Cohorts." In: *Nature Medicine* (2019). DOI: 10.1038/s41591-019-0457-8.

[48] Jerome H. Friedman. "Stochastic Gradient Boosting." In: *Computational Statistics & Data Analysis* (2002). DOI: 10.1016/S0167-9473(01)00065-2.

[49] Hong Gao et al. "The Landscape of Tolerated Genetic Variation in Humans and Primates." In: *Science* (2023). DOI: 10.1126/science.abn8197.

[50] Laura H. Goetz and Nicholas J. Schork. "Personalized Medicine: Motivation, Challenges, and Progress." In: *Fertility and Sterility* (2018). DOI: 10.1016/j.fertnstert.2018.05.006.

[51] Tiansu Gong and Dongbo Bu. "Language Models Enable Zero-Shot Prediction of RNA Secondary Structures Including Pseudoknots." In: *bioRxiv* (2024). DOI: 10.1101/2024.01.27.577533. preprint.

[52] Gráinne S. Gorman et al. "Mitochondrial Diseases." In: *Nature Reviews Disease Primers* (2016). DOI: 10.1038/nrdp.2016.80.

[53] GTEx Consortium. "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues." In: *Science* (2020). DOI: 10.1126/science.aaz1776.

[54] GTEx Portal team. *GTEx Portal*. URL: https://www.gtexportal.org/home/ (visited on 02/10/2024).

[55] Peiyong Guan and Wing-Kin Sung. "Structural Variation Detection Using Next-Generation Sequencing Data: A Comparative Technical Review." In: *Methods* (2016). DOI: 10.1016/j.ymeth.2016.01.020.

[56] Vanja Haberle and Alexander Stark. "Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation." In: *Nature Reviews Molecular Cell Biology* (2018). DOI: 10.1038/s41580-018-0028-8.

[57] Douglas Hanahan and Robert A. Weinberg. "Hallmarks of Cancer: The Next Generation." In: *Cell* (2011). DOI: 10.1016/j.cell.2011.02.013.

[58] Orla Hardiman et al. "Amyotrophic Lateral Sclerosis." In: *Nature Reviews Disease Primers* (2017). DOI: 10.1038/nrdp.2017.71.

[59] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "Boosting and Additive Trees." In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Ed. by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Springer Series in Statistics. New York, NY: Springer, 2009, pp. 337–387. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7_10.

[60] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2. Springer, 2009. ISBN: 978-0-387-84857-0.

[61] Florian R. Hölzlwimmer et al. "Aberrant Expression Prediction across Human Tissues." In: *bioRxiv* (2023). DOI: 10.1101/2023.12.04.569414. preprint.

[62] Florian Rupert Hölzlwimmer. *Aberrant Gene Expression Prediction Benchmark Based on GTEx V8*. Zenodo, Nov. 21, 2023. DOI: 10.5281/zenodo.8427311. URL: https://doi.org/10.5281/zenodo.8427311.

[63] Taishan Hu et al. "Next-Generation Sequencing Technologies: An Overview." In: *Human Immunology* (2021). DOI: 10.1016/j.humimm.2021.02.012.

[64] Catherine J. E. Ingram et al. "Lactose Digestion and the Evolutionary Genetics of Lactase Persistence." In: *Human Genetics* (2009). DOI: 10.1007/s00439-008-0593-6.

[65] INSERM US14. *Orphanet Knowledge Base Release of July 2023*. Dec. 4, 2023. URL: https://www.orphadata.com/epidemiology/ (visited on 01/05/2024).

[66] Kishore Jaganathan et al. "Predicting Splicing from Primary Sequence with Deep Learning." In: *Cell* (2019). DOI: 10.1016/j.cell.2018.12.015.

[67] Jinmeng Jia and Tieliu Shi. "Towards Efficiency in Rare Disease Research: What Is Distinctive and Important?" In: *Science China. Life Sciences* (2017). DOI: 10.1007/s11427-017-9099-3.

[68] Minna U. Kaikkonen, Michael T.Y. Lam, and Christopher K. Glass. "Non-Coding RNAs as Regulators of Gene Expression and Epigenetics." In: *Cardiovascular Research* (2011). DOI: 10.1093/cvr/cvr097.

[69] Konrad J. Karczewski et al. "Systematic Single-Variant and Gene-Based Association Testing of Thousands of Phenotypes in 426,370 UK Biobank Exomes." In: *medRxiv* (2022). DOI: 10.1101/2021.06.19.21259117. preprint.

[70] Konrad J. Karczewski et al. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." In: *Nature* (2020). DOI: 10.1038/s41586-020-2308-7.

[71] Max Karlsson et al. "A Single–Cell Type Transcriptomics Map of Human Tissues." In: *Science Advances* (2021). DOI: 10.1126/sciadv.abh2169.

[72] Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. "Current Sequence-Based Models Capture Gene Expression Determinants in Promoters but Mostly Ignore Distal Enhancers." In: *Genome Biology* (2023). DOI: 10.1186/s13059-023-02899-9.

[73] Katherine Chao. *gnomAD v4.0*. gnomAD browser. Nov. 1, 2023. URL: https://gnomad.broadinstitute.org/news/2023-11-gnomad-v4-0/ (visited on 01/30/2024).

[74] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[75] M. Kimura. "The Neutral Theory of Molecular Evolution: A Review of Recent Evidence." In: *Idengaku Zasshi* (1991). DOI: 10.1266/jjg.66.367.

[76] Martin Kircher and Kerstin U. Ludwig. "Systematic Assays and Resources for the Functional Annotation of Non-Coding Variants." In: *Medizinische Genetik* (2022). DOI: 10.1515/medgen-2022-2161.

[77] Martin Kircher et al. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." In: *Nature Genetics* (2014). DOI: 10.1038/ng.2892.

[78] Philip Kleinert and Martin Kircher. "A Framework to Score the Effects of Structural Variants in Health and Disease." In: *Genome Research* (2022). DOI: 10.1101/gr.275995.121.

[79] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. "Chromatin Accessibility and the Regulatory Epigenome." In: *Nature Reviews Genetics* (2019). DOI: 10.1038/s41576-018-0089-8.

[80] Robert Kopajtich et al. "Integration of Proteomics with Genomics and Transcriptomics Increases the Diagnostic Rate of Mendelian Disorders." In: *medRxiv* (2021). DOI: 10.1101/2021.03.09.21253187. preprint.

[81] Shunichi Kosugi et al. "Comprehensive Evaluation of Structural Variation Detection Algorithms for Whole Genome Sequencing." In: *Genome Biology* (2019). DOI: 10.1186/s13059-019-1720-5.

[82] Laura S. Kremer et al. "Genetic Diagnosis of Mendelian Disorders via RNA Sequencing." In: *Nature Communications* (2017). DOI: 10.1038/ncomms15824.

[83] Samuel A. Lambert et al. "The Human Transcription Factors." In: *Cell* (2018). DOI: 10.1016/j.cell.2018.01.029.

[84] Samuel A. Lambert et al. "The Polygenic Score Catalog as an Open Database for Reproducibility and Systematic Evaluation." In: *Nature Genetics* (2021). DOI: 10.1038/s41588-021-00783-5.

[85] Kirk Lamoreaux et al. *The Power of Being Counted – RARE-X*. June 30, 2022. URL: https://rare-x.org/case-studies/the-power-of-being-counted/ (visited on 01/03/2024).

[86] Melissa J. Landrum et al. "ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype." In: *Nucleic Acids Research* (2014). DOI: 10.1093/nar/gkt1113.

[87] Ming Ta Michael Lee and Teri E. Klein. "Pharmacogenetics of Warfarin: Challenges and Opportunities." In: *Journal of Human Genetics* (2013). DOI: 10.1038/jhg.2013.40.

[88] Seunggeung Lee et al. "Rare-Variant Association Analysis: Study Designs and Statistical Tests." In: *American Journal of Human Genetics* (2014). DOI: 10.1016/j.ajhg.2014.06.009.

[89] Binglan Li et al. "Evaluation of PrediXcan for Prioritizing GWAS Associations and Predicting Gene Expression." In: *Biocomputing 2018*. WORLD SCIENTIFIC, 2017, pp. 448–459. ISBN: 978-981-323-552-6. DOI: 10.1142/9789813235533_0041.

[90] Jian-Rong Li et al. "Genetic Variants Associated mRNA Stability in Lung." In: *BMC Genomics* (2022). DOI: 10.1186/s12864-022-08405-y.

[91] Shuwei Li et al. "Whole-Genome Sequencing of Half-a-Million UK Biobank Participants." In: *medRxiv* (2023). DOI: 10.1101/2023.12.06.23299426. preprint.

[92] Xin Li et al. "The Impact of Rare Variation on Gene Expression across Tissues." In: *bioRxiv* (2016). DOI: 10.1101/074443. preprint.

[93] Xin Li et al. "The Impact of Rare Variation on Gene Expression across Tissues." In: *Nature* (2017). DOI: 10.1038/nature24267.

[94] Rik G. H. Lindeboom, Fran Supek, and Ben Lehner. "The Rules and Impact of Nonsense-Mediated mRNA Decay in Human Cancers." In: *Nature Genetics* (2016). DOI: 10.1038/ng.3664.

[95] Rik G. H. Lindeboom et al. "The Impact of Nonsense-Mediated mRNA Decay on Genetic Disease, Gene Editing and Cancer Immunotherapy." In: *Nature Genetics* (2019). DOI: 10.1038/s41588-019-0517-5.

[96] Johannes Linder et al. "Deciphering the Impact of Genetic Variation on Human Polyadenylation Using APARENT2." In: *Genome Biology* (2022). DOI: 10.1186/s13059-022-02799-4.

[97] Johannes Linder et al. "Predicting RNA-seq Coverage from DNA Sequence as a Unifying Model of Gene Regulation." In: *bioRxiv* (2023). DOI: 10.1101/2023.08.30.555582. preprint.

[98] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. "On the Dependency of Cellular Protein Levels on mRNA Abundance." In: *Cell* (2016). DOI: 10.1016/j.cell.2016.03.014.

[99] Holli A. Loomans-Kropp and Asad Umar. "Cancer Prevention and Screening: The next Step in the Era of Precision Medicine." In: *npj Precision Oncology* (2019). DOI: 10.1038/s41698-018-0075-9.

[100] Sebastian Lunke et al. "Integrated Multi-Omics for Rapid Rare Disease Diagnosis on a National Scale." In: *Nature Medicine* (2023). DOI: 10.1038/s41591-023-02401-9.

[101] Daniel G. MacArthur et al. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." In: *Science (New York, N.Y.)* (2012). DOI: 10.1126/science.1215040.

[102] Teri A. Manolio et al. "Finding the Missing Heritability of Complex Diseases." In: *Nature* (2009). DOI: 10.1038/nature08494.

[103] Nana Matoba et al. "GWAS of 165,084 Japanese Individuals Identified Nine Loci Associated with Dietary Habits." In: *Nature Human Behaviour* (2020). DOI: 10.1038/s41562-019-0805-1.

[104] William McLaren et al. "The Ensembl Variant Effect Predictor." In: *Genome Biology* (2016). DOI: 10.1186/s13059-016-0974-4.

[105] Christian Mertes et al. "Detection of Aberrant Splicing Events in RNA-seq Data Using FRASER." In: *Nature Communications* (2021). DOI: 10.1038/s41467-020-20573-7.

[106] Remo Monti et al. "Identifying Interpretable Gene-Biomarker Associations with Functionally Informed Kernel-Based Tests in 190,000 Exomes." In: (2021). DOI: 10.1101/2021.05.27.444972.

[107] Joannella Morales et al. "A Joint NCBI and EMBL-EBI Transcript Set for Clinical Genomics and Research." In: *Nature* (2022). DOI: 10.1038/s41586-022-04558-8.

[108] Mouagip. *Aminoacid Table*. Feb. 18, 2009. URL: https://commons.wikimedia.org/wiki/File:Aminoacids_table.svg (visited on 02/04/2024).

[109] Lambert Moyon et al. "Classification of Non-Coding Variants with High Pathogenic Impact." In: *PLOS Genetics* (2022). DOI: 10.1371/journal.pgen.1010191.

[110] David R. Murdock. "Enhancing Diagnosis Through RNA Sequencing." In: *Clinics in Laboratory Medicine* (2020). DOI: 10.1016/j.cll.2020.02.001.

[111] Alex V. Nesta, Denisse Tafur, and Christine R. Beck. "Hotspots of Human Mutation." In: *Trends in Genetics* (2021). DOI: 10.1016/j.tig.2020.10.003.

[112] Sergey Nurk et al. "The Complete Sequence of a Human Genome." In: *Science (New York, N.Y.)* (2022). DOI: 10.1126/science.abj6987.

[113] Ryan P. Owen et al. "VKORC1 Pharmacogenomics Summary." In: *Pharmacogenetics and Genomics* (2010). DOI: 10.1097/FPC.0b013e32833433b6.

[114] Pan-UKB team. *Pan-Ancestry Genetic Analysis of the UK Biobank*. 2020. URL: https://pan.ukbb.broadinstitute.org.

[115] Baoxu Pang et al. "Identification of Non-Coding Silencer Elements and Their Regulation of Gene Expression." In: *Nature Reviews Molecular Cell Biology* (2023). DOI: 10.1038/s41580-022-00549-9.

[116] Karl Pearson. "X. On the Criterion That a given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling." In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (1900). DOI: 10.1080/14786440009463897.

[117] David Pellerin et al. "Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia." In: *New England Journal of Medicine* (2023). DOI: 10.1056/NEJMoa2207406.

[118] John Preskill. "Stephen Hawking (1942–2018)." In: *Science* (2018). DOI: 10.1126/science.aat6775.

[119] Florian Privé et al. "Portability of 245 Polygenic Scores When Derived from the UK Biobank and Applied to 9 Ancestry Groups from the Same Cohort." In: *The American Journal of Human Genetics* (2022). DOI: 10.1016/j.ajhg.2021.11.008.

[120] Philipp Rentzsch et al. "CADD-Splice—Improving Genome-Wide Variant Effect Prediction Using Deep Learning-Derived Splice Scores." In: *Genome Medicine* (2021). DOI: 10.1186/s13073-021-00835-9.

[121] Philipp Rentzsch et al. "CADD: Predicting the Deleteriousness of Variants throughout the Human Genome." In: *Nucleic Acids Research* (2019). DOI: 10.1093/nar/gky1016.

[122] Alejandro Reyes and Wolfgang Huber. "Alternative Start and Termination Sites of Transcription Drive Most Transcript Isoform Differences across Human Tissues." In: *Nucleic Acids Research* (2018). DOI: 10.1093/nar/gkx1165.

[123] William F. Richter et al. "The Mediator Complex as a Master Regulator of Transcription by RNA Polymerase II." In: *Nature Reviews Molecular Cell Biology* (2022). DOI: 10.1038/s41580-022-00498-3.

[124] Edin Salkovic et al. "OutSingle: A Novel Method of Detecting and Injecting Outliers in RNA-Seq Count Data Using the Optimal Hard Threshold for Singular Values." In: *Bioinformatics* (2023). DOI: 10.1093/bioinformatics/btad142.

[125] Anthony P. Schuller and Rachel Green. "Roadblocks and Resolutions in Eukaryotic Translation." In: *Nature Reviews Molecular Cell Biology* (2018). DOI: 10.1038/s41580-018-0011-4.

[126] Alexandre Segers et al. "Juggling Offsets Unlocks RNA-seq Tools for Fast Scalable Differential Usage, Aberrant Splicing and Expression Analyses." In: *bioRxiv* (2023). DOI: 10.1101/2023.06.29.547014. preprint.

[127] Dror Sharon, Adva Kimchi, and Carlo Rivolta. "OR2W3 Sequence Variants Are Unlikely to Cause Inherited Retinal Diseases." In: *Ophthalmic Genetics* (2016). DOI: 10.3109/13816810.2015.1081252.

[128] Shaun Purcell and Christopher Chang. *PLINK 1.9*. Version v1.90b6.21. Oct. 19, 2020. URL: www.cog-genomics.org/plink/1.9/.

[129] Shaun Purcell and Christopher Chang. *PLINK 2.0*. Version v2.00a3.5LM. Sept. 8, 2022. URL: www.cog-genomics.org/plink/2.0/.

[130] Jay Shendure and Hanlee Ji. "Next-Generation DNA Sequencing." In: *Nature Biotechnology* (2008). DOI: 10.1038/nbt1486.

[131] Tarjinder Singh et al. "Rare Coding Variants in 10 Genes Confer Substantial Risk for Schizophrenia." In: *Nature* (2022). DOI: 10.1038/s41586-022-04556-w.

[132] Craig Smail et al. "Integration of Rare Expression Outlier-Associated Variants Improves Polygenic Risk Prediction." In: *The American Journal of Human Genetics* (2022). DOI: 10.1016/j.ajhg.2022.04.015.

[133] Elliot Sollis et al. "The NHGRI-EBI GWAS Catalog: Knowledgebase and Deposition Resource." In: *Nucleic Acids Research* (2023). DOI: 10.1093/nar/gkac1010.

[134] Kimberly Splinter et al. "Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease." In: *New England Journal of Medicine* (2018). DOI: 10.1056/NEJMoa1714458.

[135] *Stats | gnomAD*. URL: https://gnomad.broadinstitute.org/stats (visited on 02/10/2024).

[136] Cathie Sudlow et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." In: *PLOS Medicine* (2015). DOI: 10.1371/journal.pmed.1001779.

[137] Patrick Sulem et al. "Genetic Determinants of Hair, Eye and Skin Pigmentation in Europeans." In: *Nature Genetics* (2007). DOI: 10.1038/ng.2007.13.

[138] Joseph D. Szustakowski et al. "Advancing Human Genetics Research and Drug Discovery through Exome Sequencing of the UK Biobank." In: *Nature Genetics* (2021). DOI: 10.1038/s41588-021-00885-0.

[139] Yosuke Tanigawa et al. "Significant Sparse Polygenic Risk Scores across 813 Traits in UK Biobank." In: *PLOS Genetics* (2022). DOI: 10.1371/journal.pgen.1010105.

[140] Robert E. Thurman et al. "The Accessible Chromatin Landscape of the Human Genome." In: *Nature* (2012). DOI: 10.1038/nature11232.

[141] Bin Tian and James L. Manley. "Alternative Polyadenylation of mRNA Precursors." In: *Nature Reviews Molecular Cell Biology* (2017). DOI: 10.1038/nrm.2016.116.

[142] Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. "The Personal and Clinical Utility of Polygenic Risk Scores." In: *Nature Reviews Genetics* (2018). DOI: 10.1038/s41576-018-0018-x.

[143] Emil Uffelmann et al. "Genome-Wide Association Studies." In: *Nature Reviews Methods Primers* (2021). DOI: 10.1038/s43586-021-00056-9.

[144] Nils Wagner et al. "Aberrant Splicing Prediction across Human Tissues." In: *Nature Genetics* (2023). DOI: 10.1038/s41588-023-01373-3.

[145] Quanli Wang et al. "Rare Variant Contribution to Human Disease in 281,104 UK Biobank Exomes." In: *Nature* (2021). DOI: 10.1038/s41586-021-03855-y.

[146] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." In: *The Annals of Mathematical Statistics* (1938). DOI: 10.1214/aoms/1177732360.

[147] Xiangyue Wu and Gary Brewer. "The Regulation of mRNA Stability in Mammalian Cells: 2.0." In: *Gene* (2012). DOI: 10.1016/j.gene.2012.03.021.

[148] Loïc Yengo et al. "A Saturated Map of Common Genetic Variants Associated with Human Height." In: *Nature* (2022). DOI: 10.1038/s41586-022-05275-y.

[149] Vicente A. Yépez et al. "Clinical Implementation of RNA Sequencing for Mendelian Disease Diagnostics." In: *Genome Medicine* (2022). DOI: 10.1186/s13073-022-01019-9.

[150] Vicente A. Yépez et al. "Detection of Aberrant Gene Expression Events in RNA Sequencing Data." In: *Nature Protocols* (2021). DOI: 10.1038/s41596-020-00462-5.

[151] Zachary Zappala and Stephen B. Montgomery. "Non-Coding Loss-of-Function Variation in Human Genomes." In: *Human heredity* (2016). DOI: 10.1159/000447453.

[152] Jialing Zhang et al. "Characterization of Cancer Genomic Heterogeneity by Next-Generation Sequencing Advances Precision Medicine in Cancer Treatment." In: *Precision Clinical Medicine* (2018). DOI: 10.1093/pcmedi/pby007.

[153] Yingdong Zhao et al. "TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository." In: *Journal of Translational Medicine* (2021). DOI: 10.1186/s12967-021-02936-w.

[154] Mu Zhu. "Recall, Precision and Average Precision." In: *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo* (2004).

[155] Hui Zou and Trevor Hastie. "Regularization and Variable Selection Via the Elastic Net." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2005). DOI: 10.1111/j.1467-9868.2005.00503.x.