UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Dzvenymyra-Marta Yarish

Predicting the molecular mechanisms of genetic variants

Master's Thesis (30 ECTS)

Supervisor: Kaur Alasoo, PhD

Tartu 2024

Predicting the molecular mechanisms of genetic variants

Abstract:

Understanding the molecular pathways through which GWAS variants affect complex traits is essential for uncovering disease mechanisms and aiding target prioritization. Traditional molQTL mapping, commonly used to assign the variant's mode of action, often yields numerous false positives and struggles with low-frequency variants. To address this issue, we investigated the use of machine learning models to predict the mode of action (MoA) of variants. We compiled a dataset consisting of two classes of molQTLs: splicing QTLs and gene expression QTLs influenced by chromatin accessibility (caQTLs). We evaluated the performance of two deep learning models, Enformer and ChromBPNet, which represent different approaches to predicting regulatory activity, on a set of fine-mapped caQTLs, with ChromBPNet proving to be more precise. We then developed the MoA model, integrating classic genomic features with predictions from single-task deep learning models. This model achieved nearly 90% accuracy in distinguishing between the two QTL classes, surpassing the 80% accuracy of a classifier based on scores from a single large-scale multi-task model. Additionally, we applied the MoA model to score QTLs from the eQTL catalogue, identified by either gene expression or Leafcutter (commonly used to identify sQTLs) methods. Our analysis indicated that MoA model predictions aligned well with gene expression QTLs, whereas most Leafcutter QTLs were not classified as sQTLs.

In summary, this work introduces an original dataset for MoA model training and evaluation, and presents a proof-of-concept MoA model that effectively classifies GWAS variants into splicing QTLs and gene expression QTLs influenced by chromatin accessibility.

Keywords:

QTL mapping, gene expression, chromatin accessibility, machine learning, deep learning

CERCS:

B110 Bioinformatics, medical informatics, biomathematics, biometrics

Geneetiliste variantide molekulaarsete mehhanismide ennustamine

Lühikokkuvõte:

Haiguste mehhanismide avastamiseks ja uute ravimisihtmärkide prioritiseerimise hõlbustamiseks on vaja paremini mõista neid molekulaarseid mehhanisme, mille kaudu geneetiliste variandid mõjutavad haiguseid ja teisi komplekstunnuseid. Tavaliselt kasutatakse variantide toimemehhanismide väljaselgitamiseks molekulaarsete kvantitatiivse tunnuse lookuste (ingl k molecular quantitative trait locus, molQTL) uuringud, mis peaksid aitama tuvastada, kas konkreetne geneetiline variant mõjutab RNA splaissimist (sQTL) või geeniekspressiooni (eQTL). Kahjuks ei suuda aga molQTL meetodid täpselt vahet teha splaissimise ja geeniekspressiooni mehhanismidel ning lisaks ei ole neil võimekust tuvastada haruldaste variantide mõju. Nende puuduste ületamiseks uurisime, kas ja kuidas oleks võimalik kasutada masinõpet variantide toimemehhanismide ennustamiseks. Esmalt koostasime me käsitsi kureeritud treeningandmestiku, milles olid kahte tüüpi molQTLid: splaissimist mõjutavad sQTLid ja läbi kromatiini avatuse geeniekspressiooni mõjutavad eQTLid. Seejärel võrdlesime kahe süvanärvivõrgumudeli (Enformer ja ChromBPNet) võimet ennustada geneetilise variandi mõju kromatiini avatusele ja leidsime, et ChromBPNet mudeli ennustused olid üldiselt täpsemad. Järgmiseks töötasime välja geneetilise variandi toimemehhanismi ennustamise mudeli, mis ühendas endas klassikalised genoomiülesed tunnused erinevate süvaõppemudelite ennustustega. See mudel saavutas sQTL ja eQTL klasside eristamisel peaaegu 90% täpsuse, ületades märgatavalt ühe suure alusmudeli skooridel põhineva klassifikaatori 80%-list täpsust.

Viimaks rakendasime toimemehhanismi ennustamise mudelit eQTL Catalogue andmebaasis olevat QTLid klassifitseerimiseks. Meie mudeli ennustused olid hästi kooskõlas geeniekspressiooni QTL-idega, kuid enamikku Leafcutteri meetodi poolt tuvastatud võimalikke splaissimise seoseid ei klassifitseeritud sQTL-ideks. Käesoleva töö käigus loodud uudne andmekogum ja esialgne masinõppemudel võimaldavad tulevikus paremini ennustada haigusseoseliste geneetiliste variantide toimemehhanisme.”

Võtmesõnad:

QTL kaardistamine, geeniekspressioon, kromatiini avatus, masinõpe, süvaõpe

CERCS:

B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Contents

1 Introduction

Any two humans are known to share 99.5% of their DNA. What is hidden in the remaining 0.5% of the genetic code is responsible for the dramatic diversity of humans on Earth. While some of the differences (variants) contribute to such factors as height, hair and eye color, some other cause diseases. Disorders resulting from mutations in a single gene are relatively well characterized, as these mutations typically alter the amino acid sequence of a specific protein and follow predictable Mendelian inheritance patterns. However, only 1% of genome codes for proteins [1]. The rest of the meaningful variance is located in the regulatory regions of the genome which control gene expression in different cell types. The language of gene regulation is both highly complex, and highly context-specific: a particular sequence may be interpreted as an important regulatory region in a liver cell, and completely ignored by a neuron [1]. Understanding the effect of a particular variant in regulatory sequence in a particular context is an open research question.

Genome-wide association studies (GWAS) are a widely used statistical approach for studying the genetic variants associated with a particular trait (phenotype). However, GWAS does not reveal the underlying molecular mechanisms driving the observed associations. Without understanding of the molecular mechanisms, it is hard to manipulate the expression of a trait of interest. One of the potential approaches to gain better understanding of the molecular processes responsible for a genetic disorder, is to conduct a (mol)QTL study in conjunction with a GWAS. Molecular quantitative trait locus (molQTL) is a general term for a variant with a genetic association for a quantitative level of a molecular trait, such as gene expression (eQTL), chromatin accessibility (caQTL) and splicing patterns (sQTL). Thus, the specific way in which the variant influences the trait on a molecular level can be thought of as this variant's mode of action.

Minikel et al. estimated that the probability of success for drug mechanisms with genetic support is 2.6 greater than those without. This increase is even more prominent in cases where the researchers are more confident about the causal gene that was associated with the GWAS trait [2]. Protein-coding sequence (missense) and splice variants are more likely to reveal target genes with high precision and could thus directly be used to prioritise drug targets. In contrast, association studies [3, 4], and perturbation experiments [5, 6] have found that variants regulating gene expression typically have an effect on multiple neighbouring genes, making them less useful for target prioritisation. Thus, one way to be more confident about causal genes is to prioritise sQTLs over eQTLs as drug targets.

A wide variety of methods can be used to measure a molecular trait. For example, total read count from RNA-seq data is commonly used to detect a change in gene expression [7], while splicing patterns can be determined by txrevise [8] or Leafcutter [9]. In practice, distinguishing sQTLs from eQTLs is challenging because those methods often find overlapping genetic associations without revealing mechanisms [7, 3, 10].

In general, due to differences in discovery [11], molQTLs typically explain only 50% of the common variants associations detected in GWAS studies [3, 12, 4]. Secondly, as GWAS studies scale up to include more than a million individuals, they will identify more low-frequency associations that cannot be captured by the limited sample sizes of current molQTL datasets. Thus, there is a pressing need for alternative strategies to characterise and understand the mode of action of GWAS signals.

Over the last decade, machine learning (ML) methods have successfully infiltrated various branches of industry and research. In genomics, ML models have been used for variant calling and annotation [13], predicting the variant effect on gene expression [14, 15, 16], splicing [17, 18], polyadenylation [19], and chromatin accessibility [20], and discovery of the sequence preferences of RNA- and DNA-binding proteins [21]. However, they are notoriously hard to validate and interpret, especially in the field of genomics, where sequencing data lack human-readable features and cues [22].

The goal of this work is to investigate the potential of machine learning for predicting the mode of action of genetic variants. Specifically, we aim to develop a model that can differentiate between eQTLs and sQTLs using sequence features. The approach involves several key steps:

1. Building a dataset specifically for variant MoA analysis.

2. Validating the latest splicing and chromatin accessibility prediction models on a manually curated set of variants

3. Developing a variant mode of action prediction model that incorporates both traditional and neural features.

This thesis is divided into 7 broad parts. In Section 2, we provide a foundational overview of genetic information, detailing the processes and mechanisms by which it is encoded and interpreted, and also discuss the application of ML techniques in genomics research. Then, in Section 3, we highlight the importance of this line of research and share a case study on the new sickle cell anaemia drug. After that, we move on to Section 4, where we describe the datasets, models, and technological tools that were used in this thesis. Section 5 presents the study's findings, including the datasets' analyses, comparisons between different models, and the evaluation of the MoA model. Next, in

Section 6, we discuss the significance of the results and their broader implications for advancing the field of genetics and machine learning and propose potential directions for future research. Section 7 concludes the thesis by reiterating the main results, their implications, and limitations.

2 Background

In this section, we will give a short overview of the fundamental genomic concepts and processes, methods used to study the associations between genetic variants and phenotypes, and their limitations. We will also describe the working principles of machine learning and its applications in genomics.

2.1 Genetic information

2.1.1 DNA

The code that is used to produce all life forms on Earth is stored in cells as DNA (deoxyribonucleic acid) molecules. DNA is composed of two complementary strands that twist into a double-helix structure. Each strand consists of nucleotides, which include one of four nucleobases: adenine (A), thymine (T), guanine (G), or cytosine (C). These bases are the fundamental components of DNA and are crucial for the complementary pairing of the strands. Humans have diploid cells, meaning each chromosome has a counterpart originating from one of the parents. Every human cell contains 23 chromosome pairs: 22 autosomal pairs and one pair of sex chromosomes. Due to this diploid arrangement, each genetic variation in an individual's genome may be present zero, one, or two times, defining the person's *genotype*. The human genome comprises about 3.2 billion base pairs of adenine-thymine (AT) and guanine-cytosine (GC) sequences across a single set of 23 chromosomes.

DNA is not stored loosely within the cell. Rather, it is intricately packaged into a compact structure known as chromatin. This packaging is essential to fit the lengthy DNA molecules into the relatively small nucleus of a cell and plays a crucial role in gene regulation and protection of the genetic material. DNA is stored within nucleosomes, which are the fundamental units of chromatin. A nucleosome is composed of a DNA segment wound around a histone protein core. Each core consists of eight histone proteins, which include two copies each of H2A, H2B, H3, and H4. Multiple nucleosomes together form a more compact structure known as *chromatin*. Chromatin is not a passive structure; it plays an active role in regulating gene expression [23]. The degree of packing can influence whether genes are accessible to the machinery that synthesizes RNA (transcription). Highly condensed chromatin, or heterochromatin, is usually transcriptionally inactive, whereas less condensed chromatin, or euchromatin, is typically active. To study chromatin accessibility and its underlying aspects, the regions of accessible chromatin have to be detected. To find the open chromatin sequences, there are genome-wide chromatin accessibility profiling methods available, such as DNASE-seq and ATAC-seq (see Section 4.2). In each cell type, these methods typically identify 100,000–200,000 open chromatin regions covering 1%–2% of the genome [24].

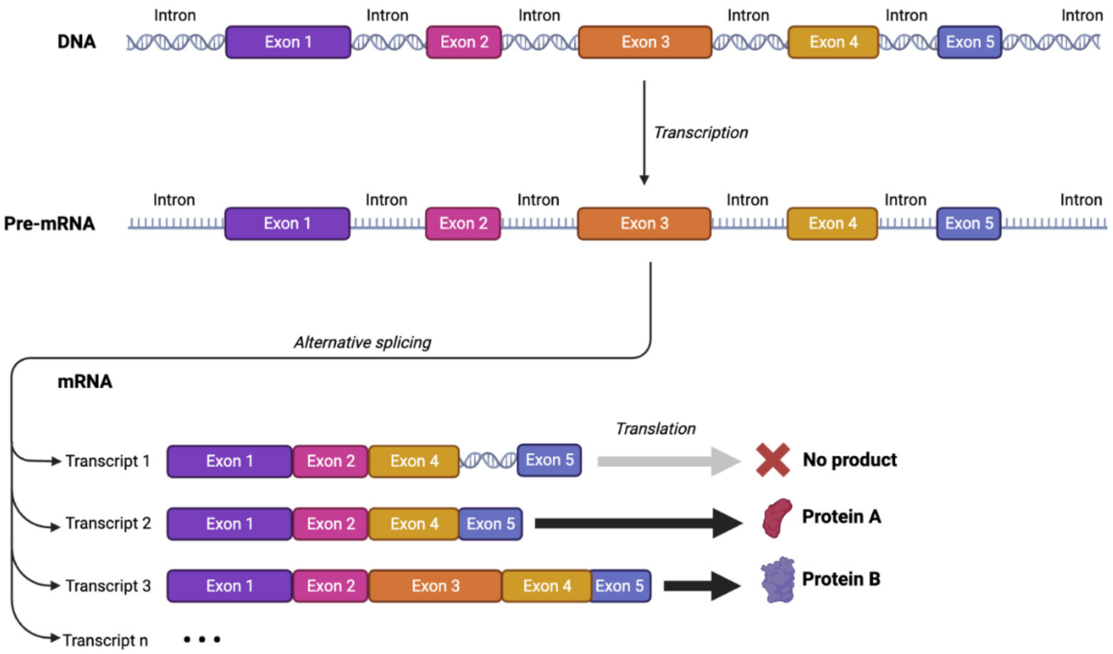


Figure 1. Gene structure and the process of alternative splicing. [25]

2.1.2 Encoding of biological information

The genome encodes two main types of information: what proteins to make (encoded in genes) and how much of each protein to make, which is referred to as *gene regulation* [1].

A gene is a segment of DNA that contains the instructions for building a functional product, such as a protein or an RNA molecule. Gene expression involves producing a corresponding protein through a two-step process. First, the information in DNA is transcribed into messenger RNA (mRNA) through transcription. During this step, RNA polymerase II synthesizes a pre-mRNA molecule using the DNA sequence as a template for complementary base-pairing. This pre-mRNA undergoes processing to become mature mRNA. The mature mRNA, a single-stranded copy of the gene, is then translated into a protein through the translation process. During *translation*, which is the second major step in gene expression, the mRNA is "read" according to the genetic code, which relates the DNA sequence to the amino acid sequence in proteins [26]. Each group of three bases in mRNA constitutes a codon, and each codon specifies a particular amino acid [26].

Another important process must occur before the mRNA can be translated into a protein - *RNA splicing*, which involves the removal or "splicing out" of certain sequences referred to as *introns*. The final mRNA thus consists of the remaining sequences, called *exons*, which are connected to one another through the splicing process [27]. One advantage of splicing is that it is possible to make different protein products from the same gene by including or excluding different combinations of exons, or by using different splice sites [1]. Figure 1 illustrates this process. This is known as alternative splicing, and the different mRNA sequences assembled from the same gene are called *transcripts*, while the different protein products are called isoforms. According to Wang et al., more than 90% of human genes undergo alternative splicing [28]. Furthermore, most alternative splicing, alternative cleavage and polyadenylation events vary between tissues, providing an important element of support for the hypothesis that alternative splicing is a principal contributor to the evolution of phenotypic complexity in mammals [28].

RNA-binding proteins (RBPs) are a diverse group of proteins that interact with RNA molecules in cells to regulate various aspects of RNA metabolism [29]. These proteins play crucial roles in RNAs' processing, transport, localization, translation, and stability. RBPs recognize specific RNA sequences or structural motifs, allowing them to bind to target RNAs selectively. RBPs can be broadly classified into several functional classes: RNA splicing factors, mRNA stability and degradation regulators, translation regulators, RNA transport and localization factors, and RNA editing factors. RNA splicing factors bind to specific sequences within the pre-mRNA, known as splice sites, to facilitate the splicing process [29].

2.1.3 Gene regulation

The process of producing specific RNAs and proteins is known as *gene expression*, and the mechanisms that control this expression are referred to as gene regulation. Gene regulation is encoded within the genome, and this regulatory information is as important as the protein-coding sequences themselves. The major focus of gene regulation is on controlling transcription. The transcription rate is controlled by core *promoter* elements and distant-acting regulatory elements such as *enhancers*, often called *cis*-regulatory elements. In eukaryotes, several proteins, called general transcription factors, recognize and bind to core promoters and form a pre-initiation complex. RNA polymerases recognize these complexes and initiate the synthesis of RNAs [30].

On top of that, processes like histone modifications and/or DNA methylation have a crucial regulatory impact on transcription. If a region is not accessible for the transcriptional machinery, e.g. in the case where the chromatin structure is compacted due to the presence of specific histone modifications, or if the promoter DNA is methylated, transcription may not start at all. Last but not least, gene activity is also controlled post-transcriptionally by non-coding RNAs such as microRNAs (miRNAs), as well as by cell signalling, resulting in protein modification or altered protein-protein interactions [31]. Those elements are a part of *trans*-regulatory machinery.

2.1.4 Genetic variation

Genetic variation refers to the differences in DNA sequences among individuals or populations. It is estimated that any two humans share approximately 99.5% of their DNA. Despite these high percentages of similarity, the absolute number of genetic differences is substantial—over 100 million genetic variants exist among humans, leading to a virtually limitless array of allele combinations (different versions of the same variant). These variations primarily fall into three categories: *single-nucleotide polymorphisms* (SNPs), insertion-deletion polymorphisms (INDELs), and structural variants (SVs). In a typical human genome, there are between 4.1 to 5.0 million deviations from the reference genome, with over 99.9% of these differences being SNPs and short INDELs [32]. The less than 0.01% remaining comprises approximately 2,100 to 2,500 SVs, which, though fewer in number, affect more bases overall—approximately 20 million bases of the sequence.

2.2 Making sense of genetic information

Genome-wide association studies (GWAS) and quantitative trait loci (QTL) analyses are two widely used techniques in genetics research [25].

2.2.1 GWAS

Genome-wide association studies (GWAS) aim to identify associations of genotypes with phenotypes by testing for differences in the allele frequency of genetic variants between individuals [33]. GWAS can consider copy-number variants or sequence variations in the human genome, although the most commonly studied genetic variants in GWAS are SNPs [33]. GWAS typically involve analysing millions of genetic variants across the entire genome, which can be time-consuming and computationally intensive [25]. However, they do not clarify the molecular mechanisms underlying the observed associations.

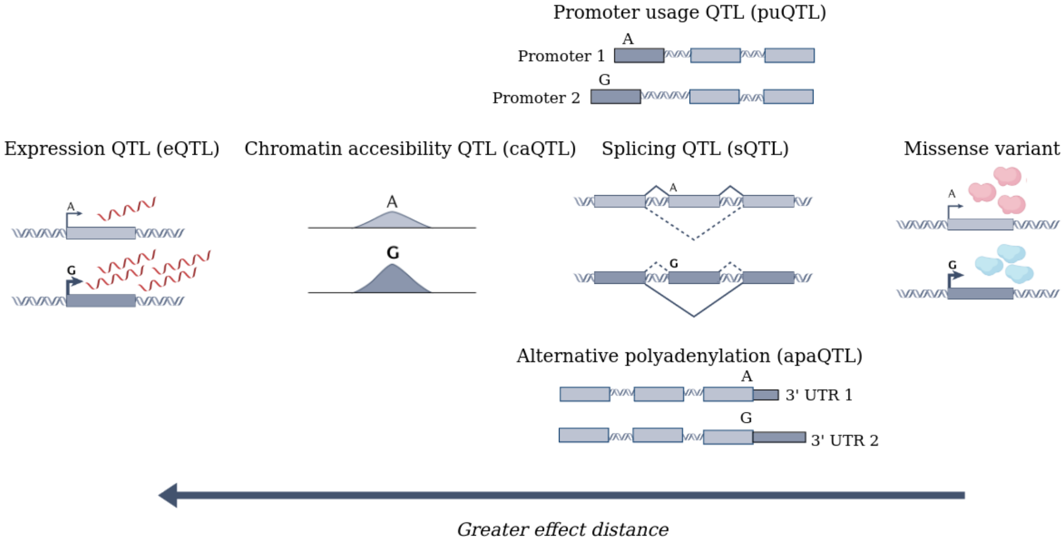


Figure 2. Modes of action of genetic variants.

Without a clear understanding of these mechanisms, it becomes challenging to intervene in the expression process of a trait of interest.

2.2.2 QTL analysis

QTL studies offer a deeper understanding of how genetic variations contribute to complex traits and diseases at a molecular level, as much of disease-associated variation is located in non-coding regions with typically unknown putative gene regulatory effects [34]. The most commonly identified QTLs include the ones for levels of gene expression (eQTLs), patterns of splicing (sQTLs), alternative polyadenylation (apaQTLs), promoter usage (puQTLs), levels of methylation of CpG sites (meQTLs), chromatin accessibility (caQTLs), levels of protein expression (pQTLs) and others. Each molQTL study typically includes association analysis for tens or hundreds of thousands of molecular features across the genome, such as all genes expressed in the studied biospecimen type [34].

QTL analysis can be performed in *-cis* and *-trans*, representing associations in physically close or distant genomic regions, respectively. In practice, this means testing for associations within a 100 kbp–1 Mbp window surrounding each studied QTL for *cis* and outside the surrounding region (>5 Mbp away or in other chromosomes) for *trans*. In this work, we focus only on *cis* associations. molQTL mapping consists of identifying statistically significant associations between genotypes and molecular phenotypes, and thus involves computing a large number of tests: for ~$1 * 10^7$ common variants and ~20,000 phenotypes, which are typical for eQTLs. This corresponds to ~$2 * 10^8$ associations for all variant–phenotype pairs in *cis* (assuming ~$1 * 10^4$ variants in each *cis* window) [34]. To make this number of computations possible, phenotypes are transformed to (approximately) follow a normal distribution (via log transform or inverse normal transform, for example), such that the associations can be calculated using linear regression. In this setup, the normalized molecular feature is treated as the dependent variable, while the genotype dosages of each participant for a specific genetic variant, along with covariates, serve as the independent variables. Because of the large number of associations tested, the perhaps most critical step in molQTL mapping is proper control of the false discovery rate (FDR), taking into account correlation between variants due to linkage disequilibrium (LD), among other potential confounding factors [34]. Linkage disequilibrium is a distinctive pattern of genetic data, when the particular alleles at nearby SNPs appear together more often than expected by chance. LD also has significant implications for genetic association studies, affecting both the power and accuracy of these analyses. Specifically, LD complicates the identification of the precise causal variants, as the observed association could be influenced by a nearby causal variant which is in LD with the variant being tested [25]. Fine-mapping methods address this issue by quantifying the probability of a variant being causal by considering the latent causal configurations that best match the observed set of effect sizes and LD between the variants [34]. In recent years, Bayesian methods have been specifically tailored for fine-mapping [35]. They compute the posterior inclusion probabilities (PIP) for each SNP as causal in a model and order the variants in decreasing order based on the PIP values. The minimal set of SNPs in the given region that captures the probable causal variant(s), known as a *credible set*, is then determined based on a pre-specified coverage probability threshold, usually 95% [25]. The ranked PIP values are summed until the cumulative probability exceeds the given threshold; the corresponding top variants are considered to form a credible set [25]. The Sum of Single Effects Model (SuSiE) is an example of such model [36]. However, fine-mapping is very sensitive to the quality of genotyping or imputation and sample size [37] and cannot be used for variants with low minor allele frequency. Besides, fine-mapping is not aware of cellular context.

Another way to think about QTL analysis is to treat the type of molecular trait mapped to a variant as its molecular mode of action. Figure 2 illustrates the incomplete set of possible modes of action, which are most relevant in the context of this work and can be determined from RNA sequencing or chromatin accessibility assays. Note that the modes of action are sorted according to the genomic distance over which they can exert an effect.

2.2.3 Genome sequencing

The above studies would not be possible without the technology for DNA sequencing, which allows to read the nucleotide sequences of DNA molecules. DNA sequencing typically can be performed in two ways - whole genome sequencing (WGS), when the entire (or almost entire) genome of an individual is read, and genotyping, which determines a person's genotype at a specific set of pre-selected SNP positions. Current commercial genotyping platforms measure between 500,000 and 2 million SNPs [1].

Besides, the unobserved genotypes can be inferred by purely computational methods via the approach called genotype imputation. This process involves comparing samples to haplotypes from a large reference panel, which includes whole genomes from a larger population. Haplotypes, which are combinations of alleles inherited together, serve as templates. Samples with missing genotypes are matched against multiple haplotypes, and the best fitting combinations are imputed into the sample, filling in the gaps in the genotype data [38]. GLIMPSE is one of the imputation methods developed to impute low-coverage sequences from reference panels using a combination of SNP enrichment and imputation methods [39].

2.3 Machine learning

Machine learning is a branch of artificial intelligence that uses statistical techniques to enable computer systems to learn from and make predictions or decisions based on data. Unlike traditional programming, where humans explicitly code all the rules and logic needed for a task, machine learning allows systems to learn these rules by identifying patterns in data. Machine learning is typically categorized into three classes: supervised learning, unsupervised learning and reinforcement learning. Supervised learning involves training a model on a labelled dataset, where the correct output (label) is provided for each input example. The model learns to map inputs to the desired output so that when it is given new examples, it can predict the corresponding outputs. The methods used with this approach include linear and logistic regression, decision trees and neural networks. In unsupervised learning, the data used to train the model is not labelled, meaning the model must find patterns and relationships within the data on its own. The goal is often to discover the underlying structure of the data, group similar data together, or reduce the number of variables. Clustering, dimensionality reduction techniques and autoencoders are used in this case. Finally, reinforcement learning is a type of machine learning where an agent learns to behave in an environment by performing actions and seeing the results.

2.3.1 Deep learning

Deep learning is a subset of machine learning that involves a class of algorithms and models known as artificial neural networks, particularly those with multiple layers or "deep" networks. These networks are designed to simulate the way human brains operate, allowing machines to process data in a complex hierarchy of layers and abstractions [40].

A typical neural network is composed of interconnected layers of nodes, or neurons. The basic architecture includes three types of layers: the input layer which receives the raw data and passes it to the next layer, hidden layers which perform various transformations on the input data, enabling the network to learn complex patterns, and the output layer that produces the final output of the network. The output layer might use a softmax activation function to generate probabilities for each class, or in regression tasks, it might produce a continuous value [41]. Convolutional Neural Networks (CNNs) are a specialized type of neural network primarily used for processing structured grid data such as images. The key components of CNNs include convolutional layers, pooling layers, and fully connected layers [42]. Convolutional layers apply convolution operations to the input data, using filters (or kernels) to detect specific features. Pooling layers reduce the spatial dimensions of the feature maps, typically using operations like max pooling or average pooling. Transformers represent a more recent advancement in deep learning, particularly in the field of natural language processing (NLP) [43]. Unlike CNNs, which are structured for spatial data, transformers excel at handling sequential data through self-attention mechanisms [44]. The self-attention mechanism allows the model to decide which elements in the sequence are the most important for a given task. By computing attention scores for each element, transformers can capture dependencies and relationships irrespective of their distance in the sequence.

Training a deep learning model involves optimizing the weights of the network using gradient descent algorithm [45] to minimize a loss function, which calculates the difference between the model's predictions and the actual targets.

2.3.2 Deep learning in genomics

Deep neural networks have proven to be sufficiently complex and versatile to capture the parameters of *cis*-regulation [46, 47, 15, 48, 49, 19, 16] as well *trans*-regulation [50, 21, 51, 20].

The transformer architecture is particularly useful for regulatory element effect prediction due to its attention mechanism, which allows each position in the DNA sequence to directly attend to all other positions [15]. This enables the model to capture long-range dependencies and integrate information from distal regulatory elements, such as enhancers, that may be located far from the transcription start site (TSS). Unlike convolutional layers, which require many successive layers to connect distant elements due to their local receptive fields, transformers can effectively increase the receptive field up to 500Kbp [16].

However, these models are generally unreliable for individual variants [52, 53], and more reliable for promoter than for enhancer variants [52]. They successfully capture major features contributing to gene expression and are valuable for many applications, but still fall short of reliably detecting all weaker effects and accurately predicting variant function. Given the complexity of cis-regulation and the large number of parameters involved, building reliable quantitative models will require much more data [54].

Moreover, to our best knowledge, there are no studies which extensively benchmark sequence-based deep learning models on predicting the effect of various molQTLs (Karollus et al. tested Enformer on predicting the impact on gene expression of GTEx eQTLs only [52]).

3 Motivation

CRISPR-Cas9 [55] genome editing is a groundbreaking technology, but genome modifications require meticulous planning and robust evidence to ensure therapeutic benefit. GWAS and molQTL analyses currently offer the most reliable targets for such precise interventions. But only last year, the first CRISPR-Cas9 drug was approved by FDA [1] for Transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD), which are both monogenic diseases caused by mutations in the haemoglobin β subunit gene (*HBB*) [56].

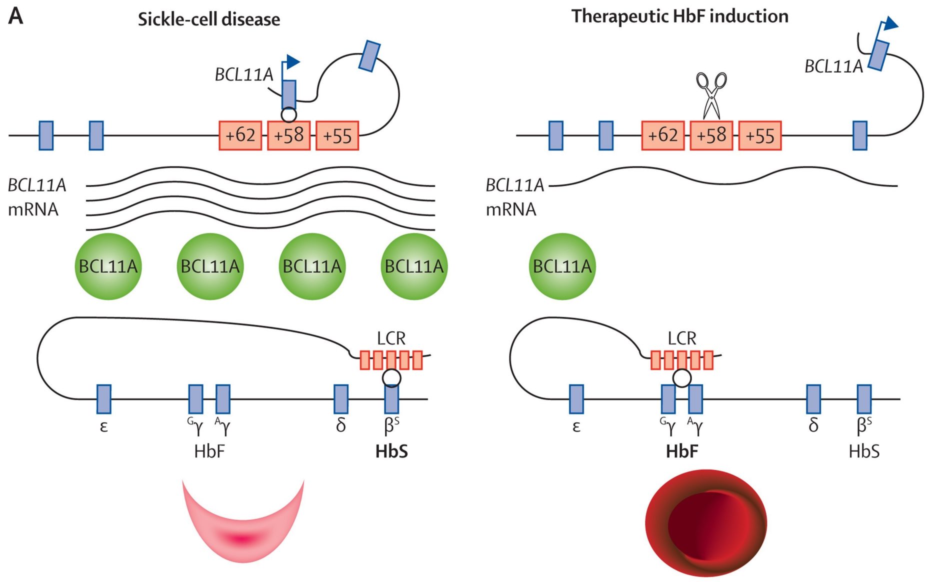


Figure 3. Therapeutic genome editing approach to induce fetal haemoglobin production in patients with sickle-cell disease [57].

Haemoglobin, the oxygen-carrying protein which is the constituent of erythrocytes (red blood cells), is a tetramer of globins. Whereas fetal haemoglobin contains two α and two γ-globins ($\alpha 2\gamma 2$), adult haemoglobin tetramer contains two α and two β-globins ($\alpha 2\beta 2$). The globin clusters undergo developmental regulation: during the latter two trimesters of gestation in human beings, fetal haemoglobin is the prevalent haemoglobin. Only after birth, in a process primarily driven by regulation of gene expression, is fetal haemoglobin replaced by adult haemoglobin [57].

[1] https://www.fda.gov/news-events/press-announcements/fda-approves-first-gene-therapies-treat-patients-sickle-cell-disease

However, the revolutionary treatment does not target the *HBB* gene directly. It implements a more complex approach. As early as 1948, there has been evidence that fetal haemoglobin can ease many of the symptoms in patients with sickle-cell disease. Then, GWAS led to the identification of the BCL11A locus as a most potent repressor of fetal haemoglobin production in human beings. *BCL11A* gene encodes a zinc-finger transcription factor. The major potential drawback to the targeting of *BCL11A* would appear to be its key functions in non-erythroid lineages. *BCL11A* has important roles in neuron development, B-cell lymphopoiesis, and dendritic cell fate, perhaps also in haemopoietic stem cells, progenitor cells and pancreatic precursors. But later, the genome editing studies have clarified that the deletion of the *BCL11A* erythroid enhancer results in the loss of *BCL11A* expression in erythroid precursors only but not in other lineages that depend on *BCL11A* such as neurons or B lymphocytes [57].

In the end, the CRISPR-Cas9 treatment deactivates the cell type-specific enhancer of *BCL11A*, which in turn stops repressing the expression of the gene coding for fetal haemoglobin [56].

This story demonstrates the importance of understanding via which mechanisms variants affect complex traits and has become a great inspiration for researchers working in this field.

4 Methods

4.1 eQTL Catalogue

eQTL Catalogue is a resource of quality-controlled, uniformly re-computed gene expression and splicing QTLs from 32 published studies [3]. It offers QTL summary statistics and fine-mapping results from the uniformly re-processed individual level eQTL data. The QTLs were identified at the level of gene expression, exon expression, transcript usage and splicing. The methods used for quantification of each molecular trait are shown in Figure 4. Gene expression was measured by counting how many reads (23 red rectangles in Figure 4) overlap the annotated exons of the gene. A similar approach was used for exon expression, except that time, this was counted for each exon separately. Transcript usage is a relative quantity - if one of the transcripts of a gene is highly used, the rest are assigned lower usage, respectively. It was estimated by looking at the

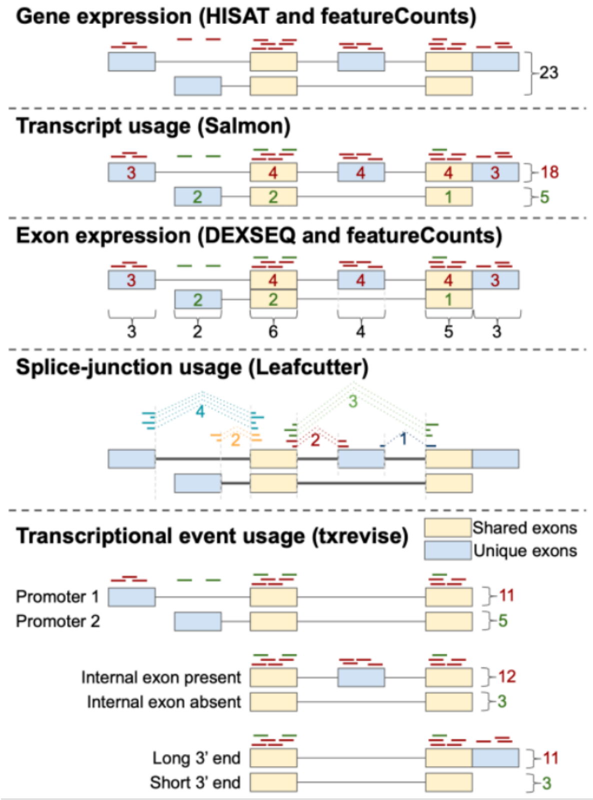

Gene expression (HISAT and featureCounts)
23
Transcript usage (Salmon)
3
4
4
4
3
18
2
2
1
5
Exon expression (DEXSEQ and featureCounts)
3
4
4
4
3
2
2
1
3
2
6
4
5
3
Splice-junction usage (Leafcutter)
4
3
2
2
1
Transcriptional event usage (txrevise)
Shared exons
Unique exons
Promoter 1
Promoter 2
11
5
Internal exon present
Internal exon absent
12
3
Long 3' end
Short 3' end
11
3

Figure 4. Overview of the five molecular trait quantification methods used in the eQTL catalogue [3].

fractions of reads that map to each transcript. Transcriptional event usage is an approach that overcomes the limitations of the previous method regarding alternative promoter and $3'$ end usage. It stratifies the reference transcripts into three events, namely promoter, splicing and $3'$ end events and then quantifies the usage separately (say something about pu and apa?). Finally, splicing events are measured by counting the number of reads that overlap exon-exon junctions.

While these methods provide a diverse set of ways to measure the molecular traits, they cannot be entirely relied upon to uncover the mode of action of the associated variants. To allow a more detailed inspection of the transcript-level associations, the authors of the eQTL catalogue developed a visualization tool in the form of static QTL coverage plots. These plots display normalised RNA-seq read coverage across all exons of the gene, exon-level QTL effect sizes and standard errors, as well as the alternative transcripts or splice junctions used in association testing [3]. The static QTL coverage plots for all 1,716,482 independent signals are available at the eQTL Catalogue Browser [2].

4.2 Data

This thesis makes use of genomic data generated by three types of biochemical assays: RNA sequencing (RNA-seq), Chromatin immunoprecipitation followed by sequencing (ChIP-seq) and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq).

RNA-seq is a powerful tool for studying gene expression and the transcriptional landscape of biological systems [25]. RNA-seq involves several key steps, beginning with mRNA extraction from a biological sample. The extracted mRNA is then processed to create a sequencing library, which involves converting the mRNA to complementary DNA (cDNA) using reverse transcription, fragmenting the cDNA into small fragments and adding adapters to the fragments [58]. Following library preparation, high-throughput sequencing technologies such as Illumina sequencing are employed to generate millions of short reads with a read depth of 10-30 million reas per sample [58]. Typically, these reads range from 50 to 500 base pairs in length and are aligned to a reference genome or transcriptome to determine the specific genomic position in which each read originated [58]. RNA-seq output is typically stored in either FASTA or FASTAQ file formats. Alignment information, in turn, is stored in a separate SAM file, which is often binarized for more effective storage and processing. The binary equivalent of SAM is called BAM. RNA-seq data can be later used to quantify various molecular traits such as gene expression, alternative splicing, and transcript usage in various biological contexts.

[2] https://elixir.ut.ee/eqtl/

ChIP-seq is a method for genome-wide profiling of DNA-binding proteins, histone modifications, or nucleosomes [59]. It relies on chromatin immunoprecipitation, a technique that isolates DNA fragments to which a specific protein or a particular class of nucleosomes is bound. After that, the enriched DNA fragments are sequenced. ChIP-seq assays can also be used to study open-chromatin regions, though less directly than ATAC-seq assays, by identifying the genomic locations of the specific histone modifications. These modifications occur on the amino acid residues of the histone tails and can influence chromatin structure and function [60]. In particular, H3K4me1 modification is typically found at active enhancers [61], and H3K27ac is highly enriched at promoter regions of transcriptionally active genes [62]. Thus, as active transcription can happen only in open chromatin regions, these epigenetic markers are an indirect way to identify them. Although, technically, ChIP-seq cannot be considered true chromatin accessibility assays.

Table 1. Overview of the chromatin QTL (cQTL) datasets included in the analysis. The number of samples is the one after MBV.

Dataset	Cell type	Assay type	Sample size	# of variants / in peaks	Genotype source
Kumasaka	LCL-1	ATAC-seq	100	1598 / 897	GLIMPSE
Kumasaka	LCL-2	ATAC-seq	91	1413 / 799	GLIMPSE
Kumasaka	LCL-3	ATAC-seq	91	1576 / 805	WGS
AFGR	LCL-4	ATAC-seq	83	1485 / 895	WGS
AFGR	LCL-5	ATAC-seq	83	1251 / 791	GLIMPSE
AFGR	LCL-6	ATAC-seq	100	1743 / 112	GLIMPSE
BLUEPRINT	naive T cells-1	ChIP H3K27ac	142	1142 / 581	WGS
BLUEPRINT	naive T cells-2	ChIP H3K4me1	103	1222 / 579	WGS
Bossini-Castillo	reg T cells-1	ChIP H3K27ac	92	365 / 167	microarray
Bossini-Castillo	reg T cells-2	ChIP H3K4me1	73	58 / 31	microarray
iPSCORE	iPSCs	ATAC-seq	64	136 / 56	microarray
BLUEPRINT	monocytes-1	ChIP H3K27ac	117	1035 / 576	WGS
BLUEPRINT	monocytes-2	ChIP H3K4me1	12	3602 / 1940	WGS

ATAC-seq (and its predecessor DNASE-seq) are techniques for studying chromatin accessibility and gene regulation landscape across different biological contexts. ATAC-seq uses a highly active Tn5 transposase to insert sequencing adapters into open regions of chromatin [63]. The resulting fragments are then sequenced and aligned to the reference genome to identify the locations of the sequenced fragments. This allows for the mapping of open chromatin regions across the genome. Next, the aligned reads are analyzed to call peaks, which represent areas of high accessibility. These peaks are then annotated by the nearest genes or regulatory elements like promoters or enhancers and can be further analyzed for motifs to identify binding sites for regulatory proteins [63].

In practice, peaks generated by ChIP-seq assays are typically several kilobases long, while ATAC-seq produces more narrow (~ 1kb) peaks.

Called peaks are typically stored in a BED (Browser Extensible Data) file format. The BED format is a simple text format consisting of one line per feature, each containing at least three columns of data: chromosome, starting position, and ending position.

4.2.1 Datasets

The overview of datasets used as sources of chromatin QTLs in different cell types is shown in Table 1. Column 'Genotype source' indicates whether the genotypes were obtained via whole genome sequencing, imputation from ATAC-seq data with GLIMPSE or microarray genotyping followed by imputation.

The list below provides more details about each of the studies. Sample sizes in Table 1 reflect the number of donors left in the dataset after the quality control with *Match BAM to VCF*(MBV) method [64], so they may differ from the ones initially reported in the publications. Chromatin QTL data re-processing and quality control were performed by Kristiina Kuningas.

1. **Kumasaka, 2018**

   In 2018, Kumasaka et al. conducted an analysis of causal interactions between regulatory elements using ATAC-seq data from 100 unrelated individuals of British ancestry [65]. The assay was performed on lymphoblastoid cell lines (LCLs). LCLs are human B cells infected by one of the most common human herpesvirus types, Epstein-Barr virus (EBV) [66]. LCLs serve as an unlimited resource of human genomic DNA, as the established cell lines apparently maintain the genome intact through generations, regardless of the viral genome persisting intracellularly [67]. Kumasaka et al. performed 75-bp paired-end sequencing in 4.4 billion sequence fragments on a HiSeq 2500 (Illumina). Whole genome sequencing data was present for 91 out of 100 individuals from the 1000 Genomes Project [3]. The genotypes for the remaining 9 samples were imputed directly from ATAC-seq data using GLIMPSE [39]. Data is available from the European Nucleotide Archive [4].

2. **African Functional Genomics Resource (AFGR)**

[3] https://www.internationalgenome.org/

[4] https://www.ebi.ac.uk/ena/browser/view/PRJEB28318

In order to lift the limitation of functional genome mapping usually being performed on European-descendent population samples only, DeGorter et al. measured gene expression using RNA sequencing in LCLs from 599 individuals from six African populations [68]. They also profiled chromatin accessibility using ATAC-Seq in a subset of 100 representative individuals from those populations. Samples were sequenced on an Illumina NextSeq using 75-bp paired-end reads. Whole genome sequencing data was present for 83 out of 100 individuals from the 1000 Genomes Project. The genotypes for the remaining 17 samples were imputed directly from ATAC-seq data using GLIMPSE [39]. The raw sequencing data is available at ENCODE [5].

3. **BLUEPRINT**

   Chen et al. performed high-resolution genetic, epigenetic, and transcriptomic profiling in three major human immune cell types (CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T cells) from up to 197 individuals [69]. Monocytes are a type of white blood cell, part of the human immune system. They circulate in the bloodstream and function as part of the body's first line of defence against pathogens [70]. Neutrophil granulocytes, commonly known as neutrophils, are essential blood cells in the innate immune and inflammatory response systems. They rapidly migrate to infection sites, usually within minutes, in response to signals from local tissue factors and resident macrophages. As a primary defence against bacterial and fungal infections, they play a critical role during the acute phase of inflammation [70]. Finally, CD4+ naive T cells are part of the adaptive immune system, representing mature helper T cells that have yet to encounter their specific antigen [70]. As a result of the study, high-resolution whole-genome sequence, RNA-seq, DNA methylation, and histone modification datasets were generated. In this thesis, we used the histone modification datasets, produced by two types of ChIP-seq assays: H3K27ac and H3K4me3 at $\geq$ 30 million reads per sample and the corresponding WGS data. BLUEPRINT dataset is a part of the eQTL catalogue, so it was also used as a source of eQTLs.

   The study was carried out as a part of the BLUEPRINT epigenome project [6].

4. **Bossini-Castillo, 2019**

   In order to identify genetic variants that control gene expression regulation in regulatory T cells isolated from healthy blood donors, Bossini-Castillo et al. profiled

[5] https://www.encodeproject.org/search/?searchTerm=AFGR&type=Experiment

[6] https://projects.ensembl.org/blueprint/

the transcriptome using RNA-seq (124 individuals), chromatin accessibility using ATAC-seq (73 individuals), promoters using H3K4me3 (88 individuals), and active enhancer and promoter regions using H3K27ac (91 individuals) [71]. Regulatory T cells (Tregs) are a specialized subpopulation of T cells that play a crucial role in maintaining immune tolerance and preventing autoimmune disease. They function primarily by suppressing the immune responses of other cells, thereby ensuring the immune system does not mistakenly attack the body's own tissues [70]. In this thesis, we used raw sequencing data produced by H3K4me3 and H3K27ac assays (initial tests suggested that the ATAC-seq data is of lower quality when compared to other ATAC-seq datasets). This dataset is a part of the eQTL catalogue, so it was also used as a source of eQTLs.

5. **iPSCORE**

   iPSCORE [7] is a collection of systematically derived and characterized iPSC lines from 222 ethnically diverse individuals. iPSCs were systematically reprogrammed from fibroblasts and analyzed for pluripotency and the presence and recurrence of somatic copy-number variants (CNVs) [72]. Induced pluripotent stem cells (iPSCs) are engineered from adult somatic cells through a process that induces a pluripotent state, enabling them to differentiate into nearly any cell type. This reprogramming is achieved by introducing specific transcription factors that revert the cells to a state resembling embryonic stem cells [70]. Germline DNA has been sequenced from blood or fibroblast samples for all 273 individuals, and other genomic data (RNA-seq, DNA methylation, ATAC-seq and genotype arrays) has been generated from the 222 iPSCs derived from a subset of these individuals. In this work, we used ATAC-seq data from 64 unrelated donors and genotypes imputed from microarray assays. This dataset is a part of the eQTL catalogue, so it was also used as a source of eQTLs.

6. **GEUVADIS**

   Lappalainen et al. performed sequencing and deep analysis of messenger RNA and microRNA from lymphoblastoid cell lines of 462 individuals (5 different populations) from the 1000 Genomes Project [73]. This dataset is a part of the eQTL catalogue, so it was used as a source of eQTLs.

7. **TwinsUK**

[7] https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000924.v4.p1

Buil et al. conducted a study using RNA-seq on ~ 400 female twin pairs (about 800 individuals) from the TwinsUK cohort to assess genetic and environmental influences on allele-specific expression. They analyzed mRNA from four tissue types — fat, skin, blood, and LCLs. The sequencing was performed using 49-bp paired-end sequencing on an Illumina HiSeq 2000 [74]. This dataset is a part of the eQTL catalogue, so it was used as a source of eQTLs.

4.2.2 ChromBPNet training data

One of the goals of this work is to benchmark the relatively novel ChromBPNet model on uniformly processed chromatin accessibility data. In pursuit of this goal, we trained six ChromBPNet models to obtain cell type specific predictions for the cell types present in caQTLs datasets. We used the ENCODE portal [8] as a source of ATAC-seq/DNASE-seq experimental data. In those experiments, .bed files with called peaks were already published, so there was no need to process any raw data. We also downloaded corresponding .bam files with alignments. Table 2 presents a more detailed description of each experiment.

Table 2. ChromBPNet training data.

Experiment	Cell type	Assay type	Data size (# of peaks)	# of replicates
ENCSR868FGK	K562 (leukemia cells)	ATAC-seq	269,718	3
ENCSR637XSC	LCLs	ATAC-seq	277,907	3
ENCSR452COS	naive CD4+ T cells	ATAC-seq	153,470	2
ENCSR159GFS	reg CD4+, CD25+ T cells	ATAC-seq	91,371	1
ENCSR485TLP	iPSCs	ATAC-seq	283,143	3
ENCSR000EPK	CD14+ monocytes	DNase-seq	88,942	1

4.2.3 MoA dataset

The Mode-of-Action (MoA) dataset was collected in three steps: manual labelling, caQTL mapping and QTLs that affect gene expression via chromatin accessibility (ceQTLs) definition.

During manual labelling, we reviewed RNA-seq QTL coverage plots and assigned them to one of six categories: eQTL, sQTL, puQTL, apaQTL, mapping bias and ambiguous. Figures 6 - 8 provide examples of plots that fall under each category.

[8] https://www.encodeproject.org/

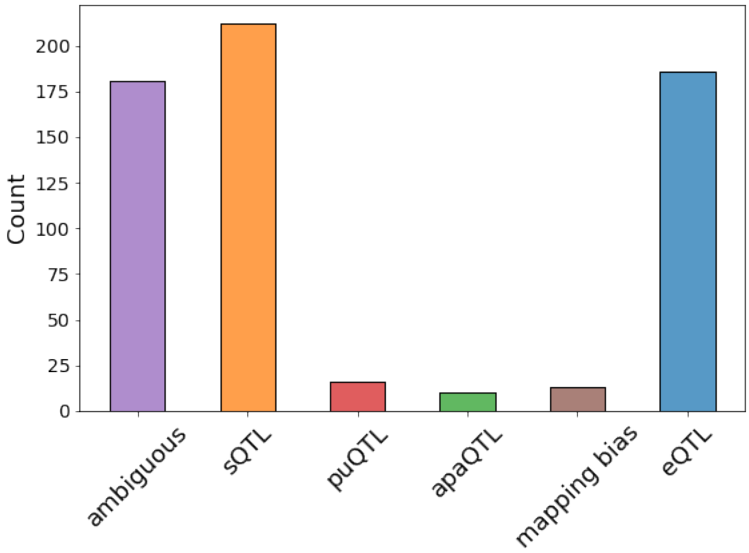


Figure 5. QTL classes distribution in the manually labelled part of the dataset.

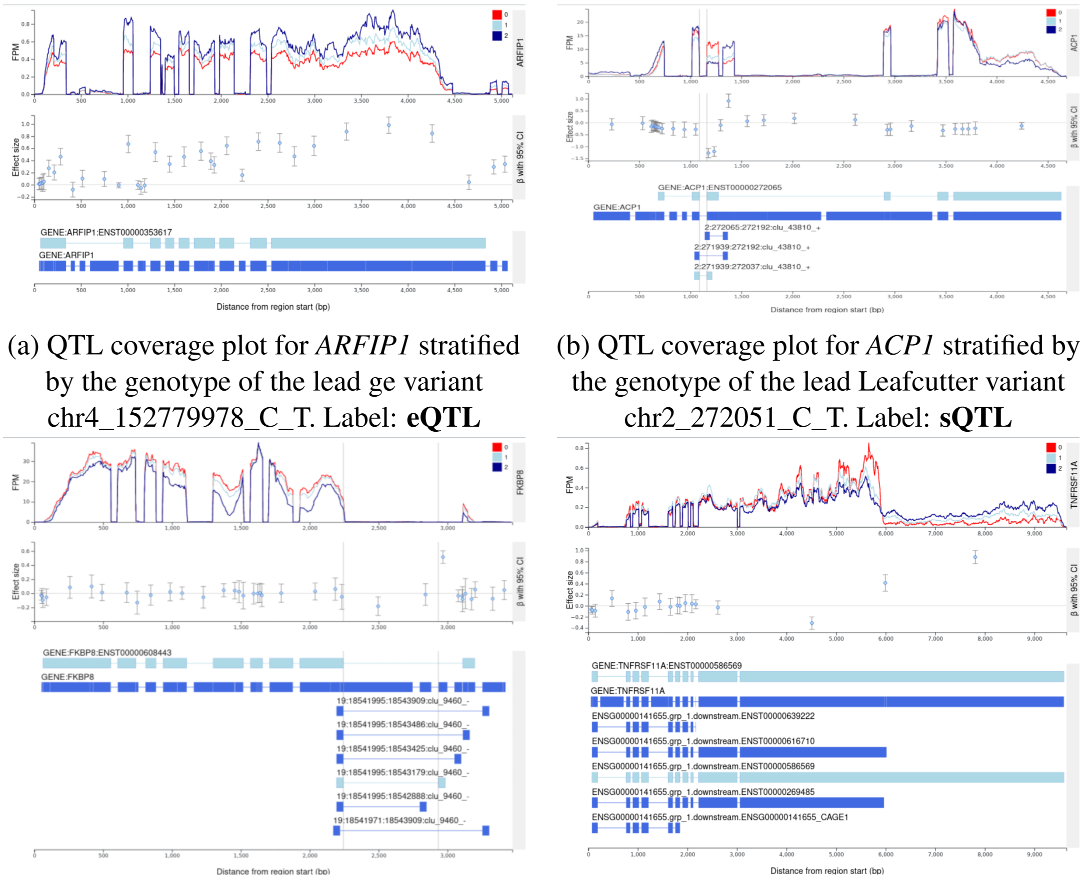
Figure 6.a shows a coverage plot with large genotype-dependent differences across all exons, which indicates a strong eQTL signal, while in Figure 6.b, the genotypes difference in coverage is most prominent for one exon only, suggesting an sQTL. Figures 6.c and 6.d illustrate coverage plots from which alternative start and end sites of transcription can be seen; therefore, these plots were assigned to puQTL and apaQTL classes respectively.

Figure7 displays a coverage plot for a QTL detected by the txrevise method. Here, the stratified by genotype difference in expression of the fourth exon is also verified by boxplots, which show that for the third genotype, the transcript with the fourth exon spliced out has the highest TMP units count - is the most expressed under this genotype.

Finally, Figure 8 demonstrates two examples of coverage plots for which the mode of action cannot be reliably identified. Specifically, Figure 8.b shows an example of mapping to the reference genome bias with the characteristic genotype-dependent bulge in read coverage in the middle of the exon.

For more coverage plot labelling examples, we refer the reader to the Supplementary materials of Kerimov et al. paper [3].

Firstly, we labelled a subset of plots associated with the GWAS traits from studies on plasma proteins [75] and metabolites (unpublished, from UK Biobank) because we assumed since the variants are associated both with GWAS and molecular traits, their effect might look more prominent on the plots. However, only 15% of the unique QTLs were classified as sQTLs. Therefore, we changed the strategy and selected all of the



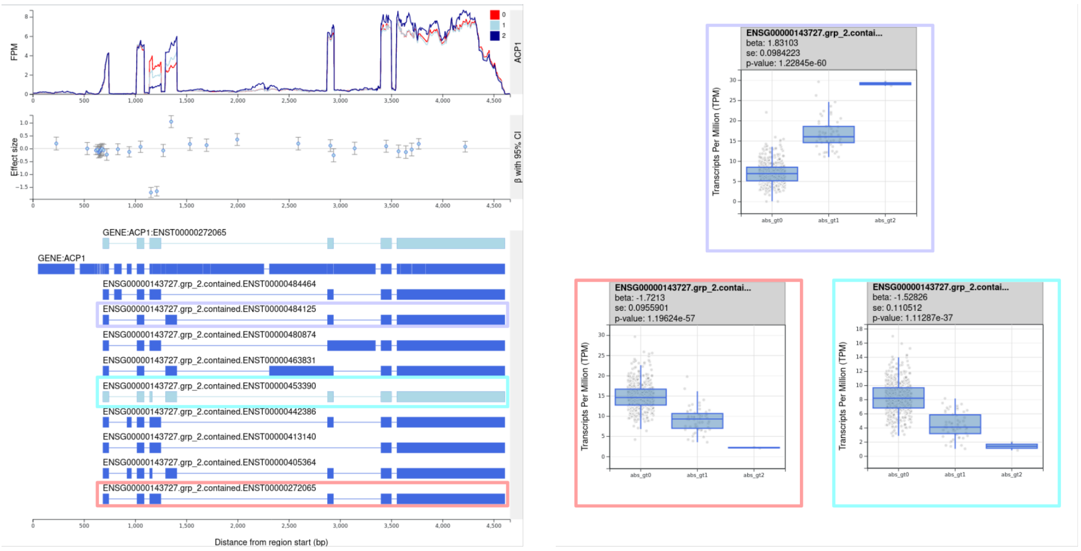
(a) QTL coverage plot for *ARFIP1* stratified by the genotype of the lead ge variant chr4_152779978_C_T. Label: **eQTL**

(b) QTL coverage plot for *ACP1* stratified by the genotype of the lead Leafcutter variant chr2_272051_C_T. Label: **sQTL**

(c) QTL coverage plot for *FKB8P* stratified by the genotype of the lead Leafcutter variant chr19_18543175_C_T.
Label: **puQTL**

(d) QTL coverage plot for *TNFRSF11A* stratified by the genotype of the lead txrevise variant chr18_62387624_A_C.
Label: **apaQTL**

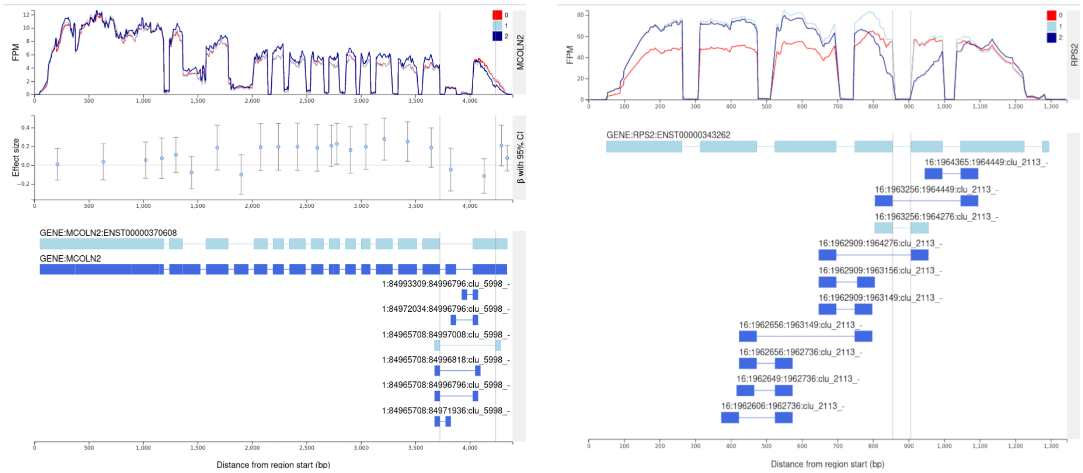
Figure 6. Example of QTL coverage plots and the mode of action labels assigned to them.



(a) QTL coverage plot for *ACP1* stratified by the genotype of the lead txrevise variant chr2_272051_C_T. Label: **sQTL**

(b) Boxplots of individual txrevise events stratified by the genotype of the lead QTL variant. The transcript with the fourth exon spliced out has the highest TPM units count in the third genotype.

Figure 7. Example of QTL coverage plot and box plots for individual txrevise events.



(a) QTL coverage plot for *MCOLN2* stratified by the genotype of the lead Leafcutter variant chr1_84997089_G_C. Label: **ambiguous**

(b) QTL coverage plot for *RPS2* stratified by the genotype of the lead Leafcutter variant chr16_1964282_A_G. Label: **mapping bias**

Figure 8. Example of QTL coverage plots for which the mode of action of the lead variant cannot be reliably determined from the plot.

variants fine-mapped with Leafcutter [9] traits from GEUVADIS and BLEUPRINT datasets with PIP ≥ 0.8 . The distribution of labels in the dataset after this step is shown in Figure 5. Even after filtering for variants enriched for high PIP and Leafcutter signal, for 28% of variants, we were not able to assign a definitive mode of action based on the coverage plots. Note that the single variant signal is usually repeated over several datasets, so the decision is based on a couple of coverage plots. Finally, we added a set of 692 disease-causing deep intronic variants affecting RNA splicing from the study by Barbosa et al. [76].

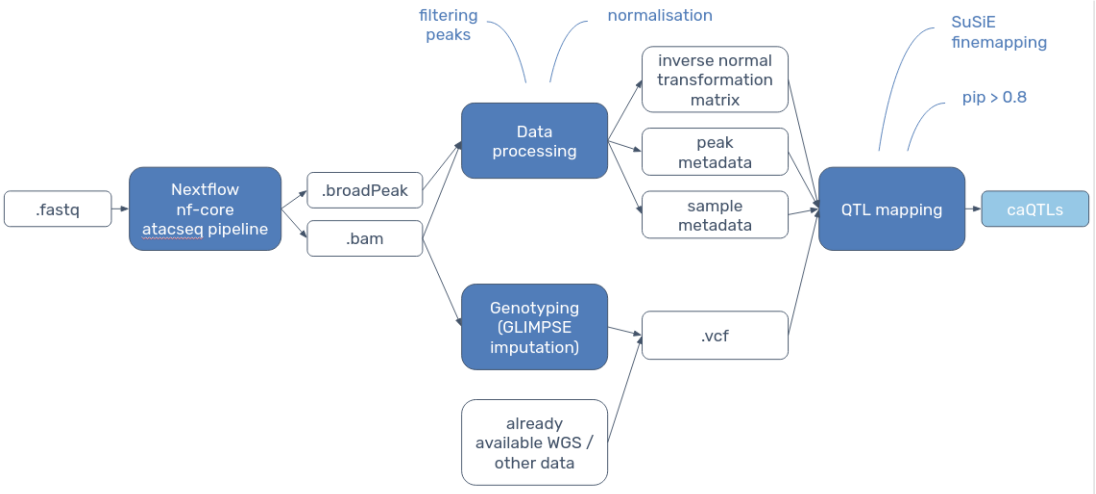


Figure 9. caQTL mapping pipeline. Credit: Kristiina Kuningas

Raw ATAC-seq/ChIP-seq data from the caQTL datasets was processed in a unified manner using the same pipeline to obtain caQTLs. The complete pipeline is shown in Figure 9. The pipeline was developed by Kristiina Kuningas.

Lastly, we selected variants fine-mapped with gene expression traits (PIP ≥ 0.8) with matching cell types as in caQTL datasets from the eQTL catalogue and computed the overlap with the caQTLs obtained in the previous step. Our underlying assumption was that if a variant is confidently fine-mapped both as an eQTL and as a caQTL, then it is likely that it affects gene expression via chromatin activity, as opposed to splicing or some other mechanisms. The results of this data manipulation are shown in Table 3. This way, we secured a set of **variants that affect gene expression by changing the chromatin structure – ceQTL**.

As a result, we collected the MoA dataset, where ceQTLs activity is mapped to a particular cell type, whereas sQTLs are considered cell-type agnostic. This decision was driven by the absence of cellular context information in the dataset curated by Barbosa et

Table 3. Variants that affect gene expression by changing the chromatin structure (ceQTL) datasets.

caQTL dataset	eQTL dataset(s)	# of caQTLs	# of eQTLs	# of shared variants / in peaks
LCLs-1	GEUVADIS, TwinsUK	1598	1876	95 / 69
LCLs-2	GEUVADIS, TwinsUK	1413	1876	89 / 70
LCLs-3	GEUVADIS, TwinsUK	1576	1876	113 / 90
LCLs-4	GEUVADIS, TwinsUK	1485	1876	69 / 57
LCLs-5	GEUVADIS, TwinsUK	1251	1876	52 / 46
LCLs-6	GEUVADIS, TwinsUK	1743	1876	67 / 60
naive T cells-1	BLUEPRINT_t-cell	1142	950	170 / 111
naive T cells-2	BLUEPRINT_t-cell	1222	950	117 / 63
reg T cells-1	Bossini-Castillo_2019	365	216	32 / 20
reg T cells-2	Bossini-Castillo_2019	58	216	8 / 7
iPSCs	iPSCORE, HipSci, PhLiPS	136	445	4 / 4
monocytes-1	BLUEPRINT_monocyte	1035	1061	170 / 110
monocytes-2	BLUEPRINT_monocyte	3602	1061	205 / 106

al. and the fact that deep learning models for splicing prediction are currently cell-type agnostic as well. While there exists evidence that the degree of sharing for sQTLs between cell types is higher than for eQTLs [7, 77], more research is required in this direction.

While initially, we planned to include four classes of QTLs in the MoA dataset, the version presented in this thesis includes only ceQTLs and sQTLs. The motivation for this decision is two-fold. Firstly, we simply were not able to collect enough samples for puQTL and apaQTL classes. Secondly, if the reader were to imagine a type of QTL as a spectrum, sQTLs and ceQTLs would be placed at opposite ends owing to the very distinct molecular mechanisms that drive these processes, making them relatively easy to tell apart. In the meantime, eQTLs are more tricky to predict and require the integration of multiple strands of evidence [3]. Furthermore, eQTLs are often defined as changing the steady-state total RNA read count [7]. But there exist multiple mechanisms which affect the total read count:

1. Enhancer variants that increase the rate of transcription, which are imperfectly proxied by caQTLs or cQTLs (chromatin QTLs) (~30-55% of all eQTLs) [78].
2. Splicing QTLs that reduce RNA stability (e.g. via NMD), resulting in reduced total read count [10].

3. Splicing QTLs that exclude a long exon, resulting in a slight decrease in total read count.

4. apaQTLs, since the 3' untranslated regions are typically long (~50% of the transcribed region), and change in their length can affect total read count. apaQTLs can also change stability by adding or removing RBP binding sites [79].

5. Other sequence variants that change stability by changing RBP binding or miRNA binding [29].

6. puQTLs that change transcript length and or stability mechanism [8].

As a result, in practice, when we want to detect an eQTL, we need to consider a whole batch of molecular mechanisms that can affect gene expression, rendering the term 'eQTL' not useful.

4.3 Models

SpliceAI [17] is a deep learning-based tool developed to predict splice sites in genomic DNA sequences using pre-mRNA sequence as input. Its architecture primarily relies on a convolutional neural network (CNN) designed to detect patterns determining where splicing occurs in the genome. SpliceAI uses dilated convolutions to capture long-range dependencies in DNA sequences effectively. This technique allows the model to have a wider receptive field, thus incorporating information from distant parts of the sequence without drastically increasing computational complexity. The authors provide SpliceAI-80nt, SpliceAI-400nt, SpliceAI-2k, and SpliceAI-10k architectures, with the features from the final convolutional layer spanning 80, 400, 2K and 10K neighbouring nucleotides, respectively. The model outputs three scores which sum to one, corresponding to the probability of the position of interest being a splice acceptor, splice donor, and neither. Then, to evaluate the splice-altering effect of a mutation, SpliceAI predicts these probabilities at each position in the pre-mRNA sequence of the gene with and without the mutation. The Δ score value for the mutation is the largest change in splice prediction scores in a window around the variant (Figure 10). The size of the sequence window is a hyperparameter, with the maximum value being equal to the input sequence length. In this work, we set it to 1000bp. For network training, authors used GENCODE-annotated pre-mRNA transcript sequences [80] on a subset of the human chromosomes and transcripts on the remaining chromosomes, with paralogs excluded, to test the network's predictions.

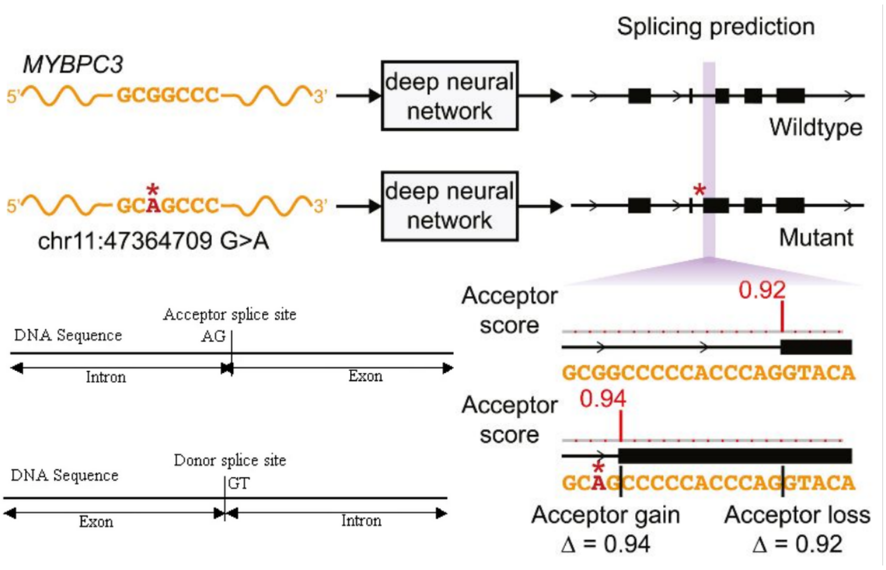


Figure 10. Schematic of using SpliceAI for accessing the splice-altering effect of a mutation [17].

Pangolin [18] is another deep-learning based tool for modelling splicing from the raw sequence. It improves over SpliceAI by predicting splicing in four tissues - heart, liver, brain and testis - separately and can predict the usage of a splice site in addition to the probability that it is spliced. Pangolin's architecture is similar to that of SpliceAI, enabling it to model features from up to 5K nucleotides both upstream and downstream of each targeted splice site. Training data for Pangolin was also collected differently. The authors processed RNA-seq data from four species (human, rhesus macaque, mouse, and rat). Then, they labelled every position within a gene body as spliced or not spliced and measured the usage of each splice site. Specifically, they marked all sites within gene bodies supported by one split read in at least 2 samples each as spliced, and all other sites as unspliced. They did not label the splice sites as donor or acceptor. Splice site usage was estimated with SpliSER [81]. We were not interested in tissue-specific predictions in this work, so we used the aggregated score.

Enformer is a Transformer-like deep learning model which predicts gene expression and chromatin states in humans and mice from DNA sequence. Using transformer layers allowed authors to significantly increase the receptive field, covering distal regulatory elements up to 100kb away while still being able to integrate their information effectively. Enformer takes as input DNA sequence of length 196,608bp and predicts 5,313 genomic tracks(transcription factors (TF) ChIP-seq, histone modification ChIP-seq, DNASE-

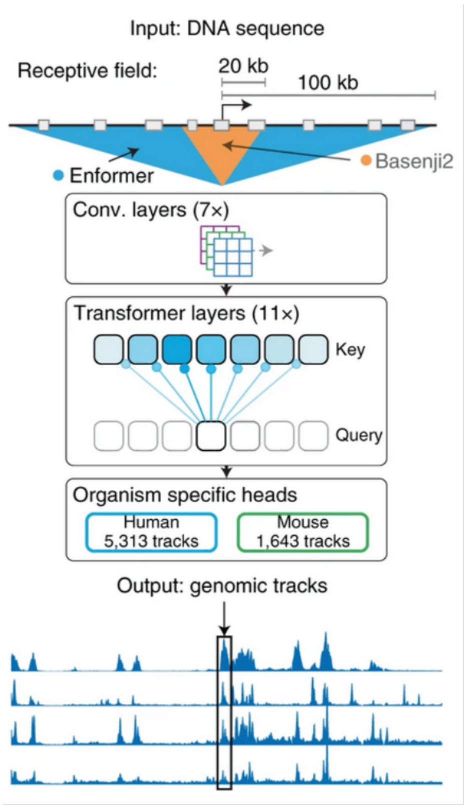


Figure 11. Enformer architecture. By using transformer architecture instead of dilated convolutions, it achieves a receptive field that detects sequence elements 100kb away [15].

seq and CAGE for gene expression) for the human genome and 1,643 tracks for the mouse genome, each of length 896 corresponding to 114,688bp aggregated into 128-bp bins (Figure 11).

ChromBPNet [20] is a fully convolutional neural network that employs dilated convolutions combined with residual connections to predict the chromatin accessibility profiles. This design allows it to have large receptive fields while using parameters efficiently. Additionally, it automatically corrects assay bias in a two-step process. Initially, it develops a simple model based on chromatin background, which accounts for enzyme effects. Subsequently, this model is used to remove the influence of the enzyme from the ATAC-seq/DNASE-seq profiles. This dual-step approach ensures that the ChromBPNet model's sequence-focused part (TF Model) does not incorporate enzymatic bias. ChromBPNet predicts the base-resolution base counts (unlike Enformer and Borzoi) and a sum of all counts in a 1000bp window from the input DNA sequence

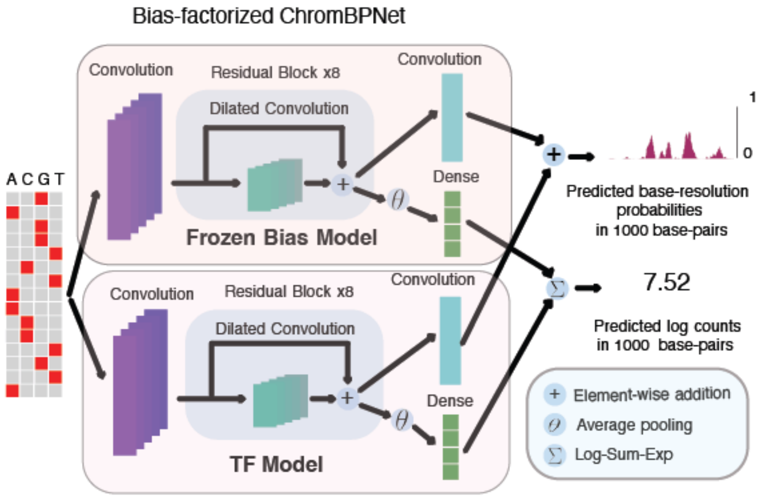


Figure 12. ChromBPNet architecture. It uses a bias model to factor out the enzymatic effect from ATAC-seq/DNASE-seq profiles [20].

of length 2114bp (Figure 12). The model requires aligned ATAC-seq/ChIP-seq reads and open chromatin peak coordinates for training. During training, it also performs sampling of negative (not accessible) regions with the same GC content (fraction of G/C bases in a sequence) as the accessible ones. Negative samples are sampled with a ratio of 0.1. A more detailed description of the data we used for training ChromBPNet models for different cell types can be found in Section 4.2.2.

Borzoi [16], Enformer's successor, is a Transformer-like deep learning model, which learns to predict cell- and tissue-specific RNA-seq coverage from DNA sequence. Borzoi uses the core Enformer architecture and employs U-net architecture with upsampling blocks to increase the output resolution to 32bp (Figure 13). For training, the authors chose to use uniformly processed RNA-seq data from the ENCODE project, which includes 900 human and 600 mouse datasets [82]. Additionally, they incorporated 2-3 replicates from each GTEx tissue, processed by the recount3 project [83]. To aid in identifying distal regulatory elements, they combined this data with thousands of training datasets from the Enformer model, featuring CAGE, DNase, ATAC, and ChIP-seq tracks.

To access the variant effect on gene expression, the 524 kb input window is centered on the SNP of interest and the model predicts coverage $y^{(\mathrm{ref})} = M(x^{(\mathrm{ref})}), y^{(\mathrm{alt})} = M(x^{(\mathrm{alt}))} \in \mathbb{R}^{16,384 \times 7,611}$ for the reference and variant patterns. Then, the L2 score is computed separately for each track across the output vector of size 16,384:

$$L2_t = \sqrt{\sum_{i=1}^{16,384} (\log(y_{i,t}^{(\mathrm{alt})} + 1) - \log(y_{i,t}^{(\mathrm{ref}} + 1)^2} \qquad (1)$$

The authors used a similar approach for predicting the splicing effect of a variant, but across the gene span only:

$$splice_t = max_{i=b_{start}}^{b_{end}} \left| \frac{y_{i,t}^{(\mathrm{alt})}}{\sum_{k=b_{start}}^{b_{end}} y_{k,t}^{(\mathrm{alt})}} - \frac{y_{i,t}^{(\mathrm{ref})}}{\sum_{k=b_{start}}^{b_{end}} y_{k,t}^{(\mathrm{ref})}} \right| \qquad (2)$$

The indices b_{start} and b_{end} in the above equation refer to the bins in y overlapping the start- and end positions of the gene span.

In our analysis, we included L2 and splicing scores calculated from all available RNA-seq tracks and a single DNASE track for each of the five cell types.

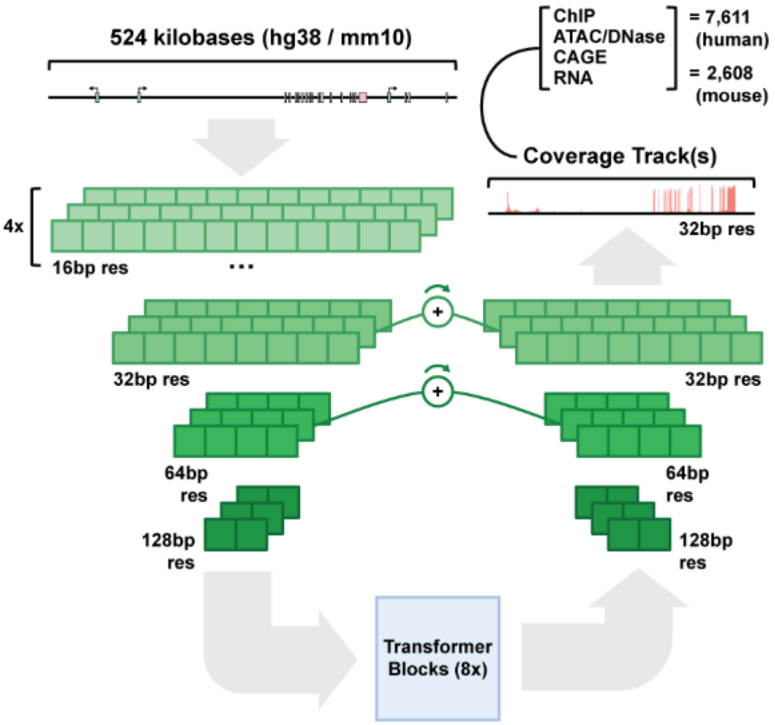


Figure 13. Borzoi model architecture. It directly predicts RNA-seq coverage and uses U-net architecture to increase the final resolution from 128bp to 32bp [16].

4.4 MoA model

MoA model is a simple binary classifier - logistic regression or decision tree - which uses a set of classic and neural features to predict the mode of action of a variant: whether it acts as a splicing QTL or gene expression QTL with effect on chromatin structure. Obviously, these two classes do not cover all of the spectrum of possible molecular mechanisms (most notably missense or loss-of-function mutations) via which the variant can affect the complex trait, so it cannot be employed as a sole decision maker for an arbitrary set of GWAS variants. However, it can be used for refining fine-mapping results or as an additional signal in a more complex analysis.

Gene expression is primarily regulated by transcription factors (TFs) that bind to DNA at promoters or enhancers [1]. This binding can be inferred from chromatin accessibility measurements, indicating potential transcription factor activity. Splicing is primarily regulated by splicing factors and RNA Binding Proteins (RBP) that bind to the transcribed pre-mRNA molecule [84]. TF binding events at enhancers are further away from the gene body and splice junctions than the binding sites of most RBPs and splicing factors. Given these distinctions, we selected the following 'classic' variant features for our model:

1. Binary variable indicating whether the variant is located within the gene body
2. Distance from the variant to the closest annotated splice junction (GENCODE v39 annotation [80])
3. Number of overlaps with open chromatin regions in 5 cell types. We used the same ENCODE DNASE/ATAC-seq experiments on which ChromBPNet models were trained.
4. Number of overlaps with binding sites of RNA binding proteins. We took the binding sites of 211 RBPs, identified by Nostrand et al. [85].

Neural features combine the predictions of three classes of deep learning models:

1. Splicing scores from SpliceAI and Pangolin. Each model produces two scores: maximum increase and decrease in the probability of a site being a splice junction in a 1000bp window around the variant.
2. Enformer SAD scores for five CAGE tracks (gene expression) and five DNASE tracks. SAD score is a difference between Enformer predictions for reference and alternative alleles, averaged over the eight flanking bins representing 1000bp window.

3. ChromBPNet difference scores for five cell types. Difference score is computed as $log_2(sum\_count_{alt} \,/\, sum\_count_{ref})$.

We used sklearn's implementation of Logistic Regression and Random Forest classifiers. For Logistic Regression, we used the L2 penalty with C=0.8. Random Forest models were fitted with the following hyperparameters: n_estimators=200, max_depth=8, min_samples_leaf=6, min_samples_split=4, max_samples=0.9.

4.5 Technology

All model training and memory-intensive data processing was run on the University of Tartu's high-performance computing centre (HPC). SpliceAI, ChromBPNet, Enformer, and Borzoi are open-source models implemented using TensorFlow [86], while Pangolin is written in PyTorch [87]. Genomic features were computed using R. The features for the MoA model can be obtained by running a single Nextflow [88] workflow, which combines together all bash scripts required to run the inference. The workflow can be found here `https://github.com/DzvinkaYarish/qtl-moa-prediction`.

5 Results

5.1 MoA dataset

Figures 14 and 15 show the MoA dataset statistics. There were no common variants between our manually labelled set and sQTLs from Barbosa et al., and only 33 of those sQTLs are discoverable via Open Targets Genetics platform [89].

0.5% (5 out of 905) of sQTLs and 26% (163 out of 624) ceQTLs localize outside of the gene body.

Ultimately, the ceQTL class was underrepresented, so we oversampled it when training the MoA model.

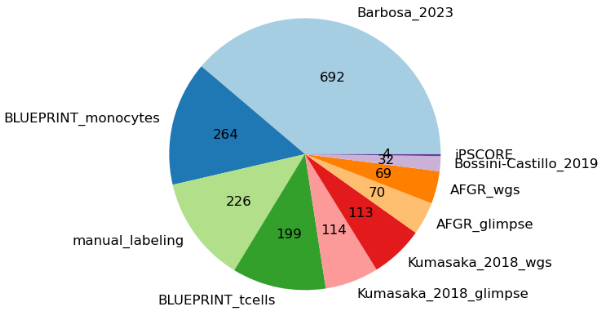


Figure 14. Sources of variants in the MoA dataset.

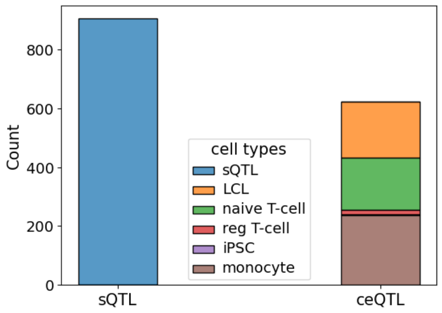


Figure 15. Classes and cell types distribution in the MoA dataset.

5.2 SpliceAI and Pangolin evaluation

Figure 16 visualizes SpliceAI and Pangolin predictions on the set of manually picked sQTLs. Out of 213 variants, both models missed the effect of 30 variants. For this evaluation, we used threshold = 0.01, while the threshold = 0.2, as suggested in the SpliceAI paper, appeared too stringent. Notably, Pangolin appears to be more conservative in its predictions than SpliceAI. Besides, models predicted a non-zero splicing effect for 17 out of 40 SNPs for which the closest annotated splice junction is outside the model's prediction window. As can be seen from the figure, splicing scores are mostly positive for those variants, so the models effectively identified novel splice junctions.

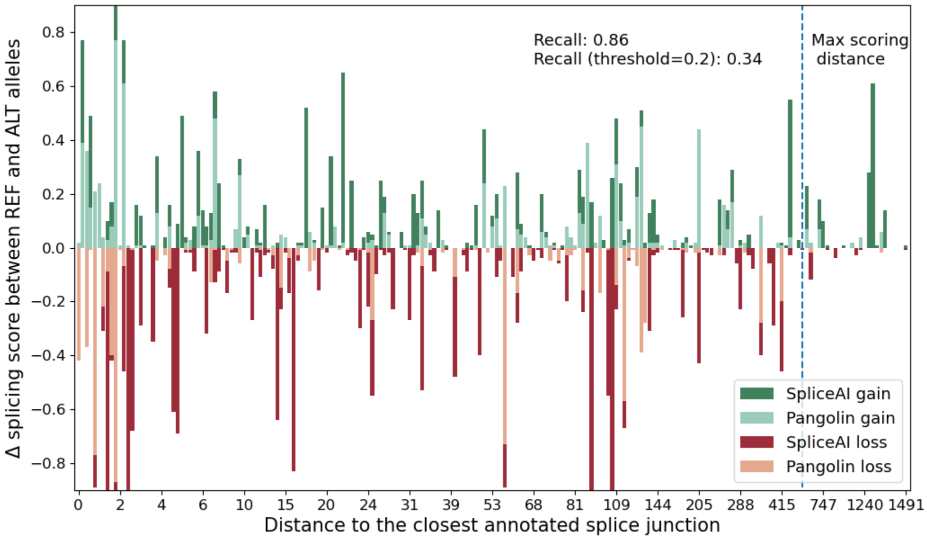


Figure 16. SpliceAI and Pangolin scores for hand-labelled set of sQTLs. Max scoring distance - length of variant left/right flanking sequences for which the scores are predicted.

5.3 ChromBPNet vs Enformer

Enformer and its successors represent a class of general-purpose (one can even say foundational) models with a large number of parameters which aspire to predict as many genomic tracks from the DNA sequence as possible. On the contrary, ChromBPNet is a specialized model which was designed with the specifics of chromatin accessibility assays in mind and is much easier and faster to train.

Table 4 compares the ChromBPNet and Enformer performances on detecting the effect size of caQTLs from uniformly processed datasets. The last 7 rows show results for subsets of caQTLs created by overlapping variants from two datasets. ChromBPNet outperforms Enformer on all datasets except for naive T cells. It might be because of the low quality of a particular ENCODE experiment with naive T cells ATAC-seq data on which ChromBPNet was trained or due to the fact that naive T cells caQTLs were detected from ChIP-seq assay, which produces broader peaks than DNASE/ATAC-seq. Secondly, we can see that the genotype source mildly affects the predictions - LCL-1,2,5,6 used genotypes imputed with GLIMPSE, while LCL-3 and LCL-4 were processed with the whole genome sequencing data.

Table 4. Performance of ChromBPNet and Enformer on caQTL datasets.

Dataset	r_s for variants in peaks			r_s for variants outside peaks		
	CBPNet x effect	Enformer x effect	CBPNet x Enformer	CBPNet x effect	Enformer x effect	CBPNet x Enformer
LCL-1	**0.727**	0.623	0.725	0.009	0.061	0.212
LCL-2	**0.735**	0.594	0.704	0.032	0.095	0.233
LCL-3	**0.778**	0.658	0.731	-0.06	0.006	0.218
LCL-4	**0.802**	0.74	0.778	0.006	0.041	0.245
LCL-5	**0.775**	0.705	0.746	0.074	0.015	0.301
LCL-6	**0.775**	0.69	0.754	0.02	0.022	0.224
naive T cells-1	0.687	**0.742**	0.74	0.028	0.136	0.163
naive T cells-2	0.669	**0.705**	0.732	0.062	0.141	0.235
reg T cells-1	0.705	**0.734**	0.816	0.107	0.056	0.14
reg T cells-2	**0.838**	0.73	0.851	0.18	-0.011	0.671
iPSCs	**0.75**	0.743	0.818	-0.03	-0.011	0.461
monocytes-1	**0.742**	0.651	0.743	0.277	0.316	0.441
monocytes-2	**0.767**	0.625	0.647	0.24	0.128	0.359
LCL-2 & 5	**0.749**	0.607	0.747	-	-	-
LCL-3 & 4	**0.78**	0.663	0.764	-	-	-
LCL-2 & 4	**0.741**	0.631	0.736	-	-	-
LCL-3 & 5	**0.798**	0.633	0.725	-	-	-
naive T cells-1 & 2	0.664	**0.706**	0.706	-	-	-
reg T cells-1 & 2	0.703	**0.805**	0.824	-	-	-
monocytes-1 & 2	**0.749**	0.652	0.738	-	-	-

Fine-mapping of chromatin accessibility associated variants with a testing window of 400Kbp resulted in roughly half of the caQTLs detected outside of the peaks (see Table 1). These results are consistent with the other caQTL studies [90, 91], so we did not exclude the variants outside of peaks from our analysis. However, the ChromBPNet authors designed the model's default architecture to have a receptive field of 1000bp, an average width of the peak detected by DNASE/ATAC-seq methods. So, we did not expect to see a non-zero effect predicted for variants located outside of peaks. Conversely, Enformer boasts a 200Kbp receptive field. But, as can be seen from the right part of Table 4, the correlation between measured and predicted effect size is low. Evidently, the simple scalar score used to evaluate the variant effect from tracks predicted by Enformer cannot capture the long-distance effect. In addition to this, no variants outside the peaks are shared between two same cell type datasets, suggesting that many of these variants could be fine-mapping false positives (i.e. variants with high posterior inclusion probabilities that are actually not causal).

Figures 17 and 18 offer a more detailed view of ChromBPNet predictions. Variants outside of peaks show almost zero effect, while variants inside peaks demonstrate high concordance with the measured effect sign. As illustrated by Figure 18, a relatively low fraction of caQTLs inside of peaks were missed by ChromBPNet, and for many variants, the effect size was overestimated.

Table 5. Spearman correlation between measured and predicted effect for different cell type ChromBPNet models for caQTL datasets.

Dataset	CBPNet model					
	monocytes	reg T cells	naive T cells	iPSCs	LCLs	leukemia cells
LCL-1	0.551	0.543	0.57	0.294	**0.708**	0.373
LCL-2	0.55	0.557	0.59	0.282	**0.711**	0.348
LCL-3	0.607	0.601	0.623	0.33	**0.755**	0.381
LCL-4	0.689	0.66	0.713	0.309	**0.792**	0.444
LCL-5	0.663	0.658	0.697	0.292	**0.77**	0.439
LCL-6	0.649	0.615	0.659	0.274	**0.771**	0.43
naive T cells-1	0.517	0.689	**0.678**	0.365	0.618	0.477
naive T cells-2	0.499	0.665	**0.658**	0.215	0.629	0.421
reg T cells-1	0.607	**0.7**	0.693	0.328	0.73	0.568
reg T cells-2	0.754	**0.825**	0.811	0.276	0.773	0.694
iPSCs	0.356	0.355	0.465	**0.737**	0.448	0.456
monocytes-1	**0.74**	0.473	0.624	0.215	0.527	0.477
monocytes-2	**0.764**	0.411	0.63	0.165	0.538	0.439

Lastly, since open chromatin regions are highly cell type specific [92], we wanted to study to what extent this specificity would be captured by ChromBPNet. Therefore, we proceeded to perform caQTLs scoring for each cell type with all six ChromBPNet models. The results are gathered in Table 5. As expected, the correct cell type model produced the highest correlation with the experimental effect size (except for naive T cells), and since iPSCs differ the most from the rest of the cell types, other cell type models fail to predict chromatin activity in them.

5.4 MoA model

We evaluated the MoA model in two stages: firstly, we performed 5-fold cross-validation on the MoA dataset and compared the metrics with the classifiers trained with Borzoi

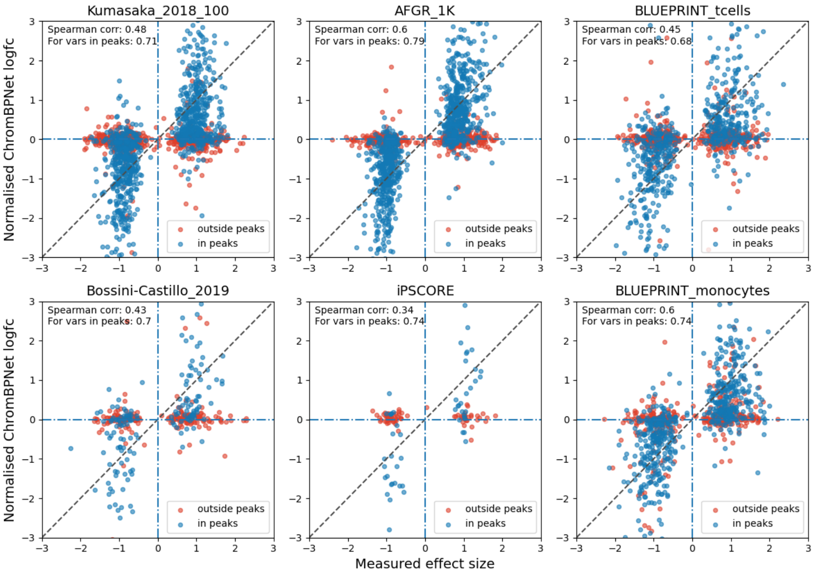


Figure 17. Predicted vs measured effect. x-axis shows beta effect values obtained from fine-mapping caQTLs, while y-axis shows normalized ChromBPNet *logfc* predictions.

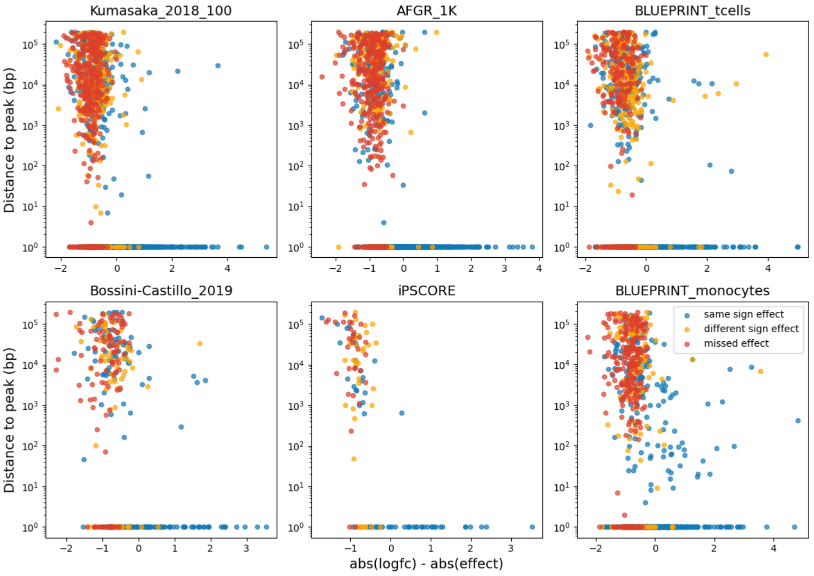


Figure 18. Error in predicted effect vs distance between the variant and the closest peak. x-axis shows the difference between absolute values of measured and predicted effect values.

scores only. After that, we scored the variants from the eQTL catalogue, quantified either by gene expression or Leafcutter method.

Table 6 presents f1 scores for logistic regression and random forest classifiers trained with different sets of features. We also trained separate models for the cell types with a sufficient number of ceQTLs. We can observe a clear improvement in accuracy after combining classic and neural features. Besides, there is no significant difference in accuracy between models which use cell type specific features and those which use all available features.

In Borzoi paper, the authors demonstrate its ability to distinguish between sQTLs, eQTLs and matched set of negatives [16]. Therefore, we decided to contrast our composite MoA model against a monolithic unified model such as Borzoi. As shown in Table 7, our MoA model with combined features from single-task neural networks consistently outperforms Borzoi-based classifiers even without the classic genomic features.

Table 6. MoA model evaluation on the MoA dataset (f1 score).

Cell type	Logistic Regression				Random Forest			
	Classic	Neural	All	All (cell type specific)	Classic	Neural	All	All (cell type specific)
All	0.675	0.848	0.846	-	0.751	0.865	0.867	-
LCL	0.702	0.814	0.832	0.847	0.771	0.864	0.86	0.88
monocytes	0.67	0.837	0.867	0.864	0.734	0.885	0.881	0.87
naive T cells	0.74	0.851	0.866	0.873	0.836	0.803	0.875	0.87

Table 7. Comparison of MoA model and Borzoi-based classifiers (f1 score)
* - cell type specific features are used.

Cell type	Logistic Regression				Random Forest			
	Borzoi	MoA (Neural)	Borzoi*	MoA (Neural)*	Borzoi	MoA (Neural)	Borzoi*	MoA (Neural)*
All	0.71	0.848	-	-	0.8	0.865	-	-
LCL	0.637	0.814	0.604	0.847	0.825	0.864	0.827	0.88
monocytes	0.614	0.837	0.588	0.864	0.819	0.885	0.774	0.87
naive T cells	0.577	0.851	0.648	0.873	0.829	0.803	0.786	0.87

Although the MoA dataset serves as a decent high-confidence benchmark, we also wanted to test the MoA model in a less controlled environment. For that purpose, we selected all QTLs from the eQTL catalogue, which were detected by gene expression

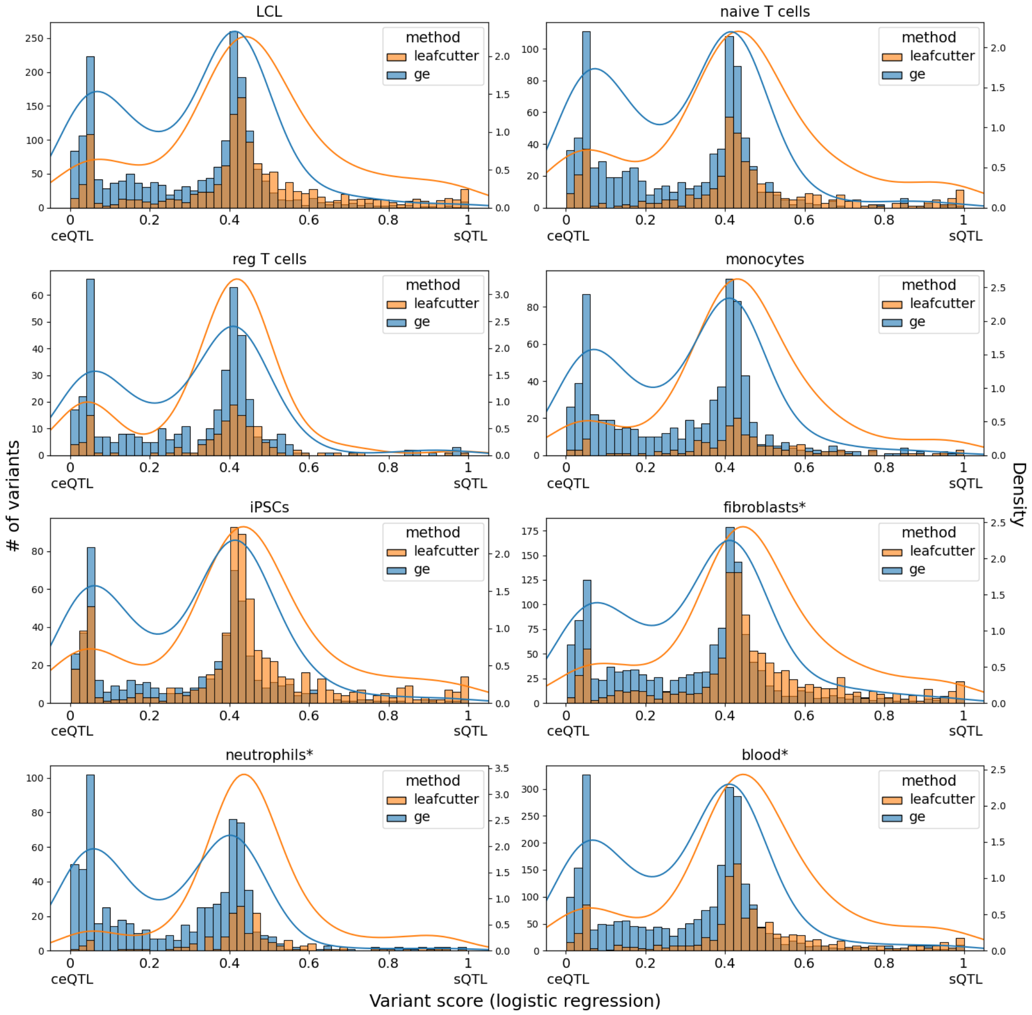


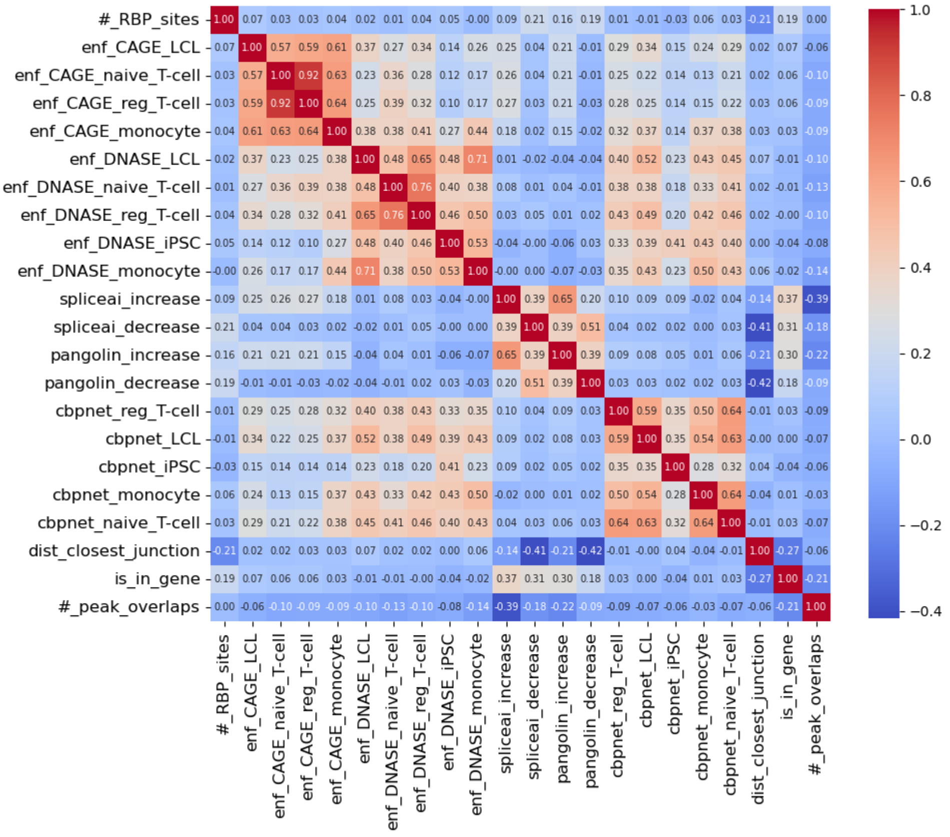
Figure 19. Distribution of the MoA model probabilities for eQTL catalogue QTLs. * - cell type/tissue not present in the MoA dataset.

(total read count) (alleged eQTLs) or Leafcutter (alleged sQTLs) methods with PIP ≥ 0.8 in five cell types present in the MoA dataset plus fibroblasts, neutrophils and blood. We then excluded those QTLs which are present in the MoA dataset. Figure 19 depicts the distribution of the MoA model probabilities of a variant being an sQTL for each cell type. While for eQTLs we can observe a bimodal distribution, shifted towards lower sQTLs probabilities (so higher ceQTL probability), this is not the case for QTLs quantified with

Leafcutter. In most cases, the model is unsure about the Leafcutter sQTls, suggesting that the classification of QTLs based on the quantification method only is unreliable.

5.5 Feature analysis

In order to understand the interactions between features and their contributions to the final model score, we conducted an extensive feature analysis. Figure 20 displays the correlation pattern between features. We can observe distinct clusters of high correlations for Enformer, ChromBPNet and SpliceAI/Pangolin predictions. Interestingly, the correlations between Enformer DNASE and ChromBPNet scores are not higher than 0.5. Regarding the classic features, number of peak overlaps is negatively correlated with the SpliceAI scores, and number of RBP sites exhibits the highest degree of independence relative to the other features analyzed.



	#_RBP_sites	enf_CAGE_LCL	enf_CAGE_naive_T-cell	enf_CAGE_reg_T-cell	enf_CAGE_monocyte	enf_DNASE_LCL	enf_DNASE_naive_T-cell	enf_DNASE_reg_T-cell	enf_DNASE_iPSC	enf_DNASE_monocyte	spliceai_increase	spliceai_decrease	pangolin_increase	pangolin_decrease	cbpnet_reg_T-cell	cbpnet_LCL	cbpnet_iPSC	cbpnet_monocyte	cbpnet_naive_T-cell	dist_closest_junction	is_in_gene	#_peak_overlaps
#_RBP_sites	1.00	0.07	0.03	0.03	0.04	0.02	0.01	0.04	0.05	-0.00	0.09	0.21	0.16	0.19	0.01	-0.01	-0.03	0.06	0.03	-0.21	0.19	0.00
enf_CAGE_LCL	0.07	1.00	0.57	0.59	0.61	0.37	0.27	0.34	0.14	0.26	0.25	0.04	0.21	-0.01	0.29	0.34	0.15	0.24	0.29	0.02	0.07	-0.06
enf_CAGE_naive_T-cell	0.03	0.57	1.00	0.92	0.63	0.23	0.36	0.28	0.12	0.17	0.26	0.04	0.21	-0.01	0.25	0.22	0.14	0.13	0.21	0.02	0.06	-0.10
enf_CAGE_reg_T-cell	0.03	0.59	0.92	1.00	0.64	0.25	0.39	0.32	0.10	0.17	0.27	0.03	0.21	-0.03	0.28	0.25	0.14	0.15	0.22	0.03	0.06	-0.09
enf_CAGE_monocyte	0.04	0.61	0.63	0.64	1.00	0.38	0.38	0.41	0.27	0.44	0.18	0.02	0.15	-0.02	0.32	0.37	0.14	0.37	0.38	0.03	0.03	-0.09
enf_DNASE_LCL	0.02	0.37	0.23	0.25	0.38	1.00	0.48	0.65	0.48	0.71	0.01	-0.02	-0.04	-0.04	0.40	0.52	0.23	0.43	0.45	0.07	-0.01	-0.10
enf_DNASE_naive_T-cell	0.01	0.27	0.36	0.39	0.38	0.48	1.00	0.76	0.40	0.38	0.08	0.01	0.04	-0.01	0.38	0.38	0.18	0.33	0.41	0.02	-0.01	-0.13
enf_DNASE_reg_T-cell	0.04	0.34	0.28	0.32	0.41	0.65	0.76	1.00	0.46	0.50	0.03	0.05	0.01	0.02	0.43	0.49	0.20	0.42	0.46	0.02	-0.00	-0.10
enf_DNASE_iPSC	0.05	0.14	0.12	0.10	0.27	0.48	0.40	0.46	1.00	0.53	-0.04	-0.00	-0.06	0.03	0.33	0.39	0.41	0.43	0.40	0.00	-0.04	-0.08
enf_DNASE_monocyte	-0.00	0.26	0.17	0.17	0.44	0.71	0.38	0.50	0.53	1.00	-0.00	0.00	-0.07	-0.03	0.35	0.43	0.23	0.50	0.43	0.06	-0.02	-0.14
spliceai_increase	0.09	0.25	0.26	0.27	0.18	0.01	0.08	0.03	-0.04	-0.00	1.00	0.39	0.65	0.20	0.10	0.09	0.09	-0.02	0.04	-0.14	0.37	-0.39
spliceai_decrease	0.21	0.04	0.04	0.03	0.02	-0.02	0.01	0.05	-0.00	0.00	0.39	1.00	0.39	0.51	0.04	0.02	0.02	0.00	0.03	-0.41	0.31	-0.18
pangolin_increase	0.16	0.21	0.21	0.21	0.15	-0.04	0.04	0.01	-0.06	-0.07	0.65	0.39	1.00	0.39	0.09	0.08	0.05	0.01	0.06	-0.21	0.30	-0.22
pangolin_decrease	0.19	-0.01	-0.01	-0.03	-0.02	-0.04	-0.01	0.02	0.03	-0.03	0.20	0.51	0.39	1.00	0.03	0.03	0.02	0.02	0.03	-0.42	0.18	-0.09
cbpnet_reg_T-cell	0.01	0.29	0.25	0.28	0.32	0.40	0.38	0.43	0.33	0.35	0.10	0.04	0.09	0.03	1.00	0.59	0.35	0.50	0.64	-0.01	0.03	-0.09
cbpnet_LCL	-0.01	0.34	0.22	0.25	0.37	0.52	0.38	0.49	0.39	0.43	0.09	0.02	0.08	0.03	0.59	1.00	0.35	0.54	0.63	-0.00	0.00	-0.07
cbpnet_iPSC	-0.03	0.15	0.14	0.14	0.14	0.23	0.18	0.20	0.41	0.23	0.09	0.02	0.05	0.02	0.35	0.35	1.00	0.28	0.32	0.04	-0.04	-0.06
cbpnet_monocyte	0.06	0.24	0.13	0.15	0.37	0.43	0.33	0.42	0.43	0.50	-0.02	0.00	0.01	0.02	0.50	0.54	0.28	1.00	0.64	-0.04	0.01	-0.03
cbpnet_naive_T-cell	0.03	0.29	0.21	0.22	0.38	0.45	0.41	0.46	0.40	0.43	0.04	0.03	0.06	0.03	0.64	0.63	0.32	0.64	1.00	-0.01	0.03	-0.07
dist_closest_junction	-0.21	0.02	0.02	0.03	0.03	0.07	0.02	0.02	0.00	0.06	-0.14	-0.41	-0.21	-0.42	-0.01	-0.00	0.04	-0.04	-0.01	1.00	-0.27	-0.06
is_in_gene	0.19	0.07	0.06	0.06	0.03	-0.01	-0.01	-0.00	-0.04	-0.02	0.37	0.31	0.30	0.18	0.03	0.00	-0.04	0.01	0.03	-0.27	1.00	-0.21
#_peak_overlaps	0.00	-0.06	-0.10	-0.09	-0.09	-0.10	-0.13	-0.10	-0.08	-0.14	-0.39	-0.18	-0.22	-0.09	-0.09	-0.07	-0.06	-0.03	-0.07	-0.06	-0.21	1.00

Figure 20. Spearman correlation between the MoA model features.

SHAP (SHapley Additive exPlanations) values [93] are derived from cooperative game theory and provide interpretability to ML models by quantifying the contribution of each feature to the prediction output. These values offer both local explanations, relevant for individual predictions, and global insights that illuminate the overall importance of features across a model.

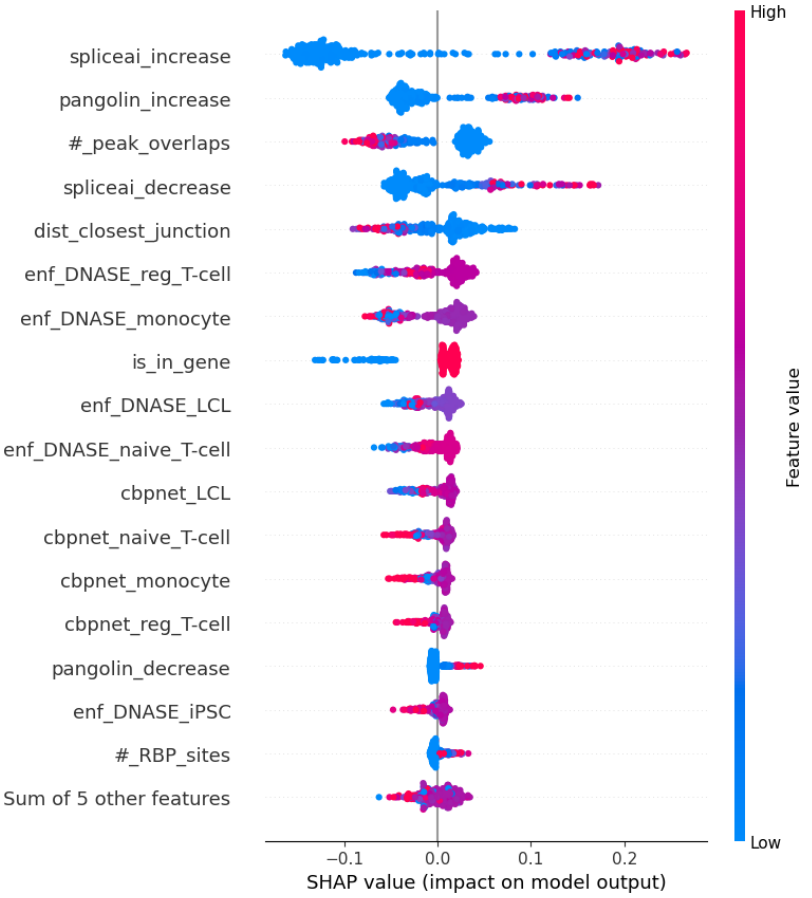


Figure 21. SHAP values plot, displaying a summary of how the top features impact the MoA model's (Random Forest) output. Each dot on each row represents one variant from the test set.

Figure 21 illustrates the top 17 features used by the Random Forest MoA model, the features are ordered by the mean absolute value of SHAP values for each feature. All features were scaled to fall into the $[0, 1]$ range. As a result, for Enformer and ChromBPNet scores, which are negative when the alternative allele makes the peak shrink and positive otherwise, scaled feature values near 0.5 indicate no effect. The plot was created with respect to the probability of a variant being a sQTL, so positive SHAP values reflect the contribution of a feature to the increase in sQTL probability, while negative values translate into the higher probability of a variant classified as ceQTL. We can immediately see that SpliceAI/Pangolin features are the most impactful, as well as the number of chromatin peak overlaps and distance to the closest splice junction. Interestingly, the decrease in the Pangolin score is far less significant than the decrease in the SpliceAI score. Additionally, Enformer DNASE features tend to have a marginally greater effect than the ChromBPNet scores, but the difference is minor. Notably, for ChromBPNet the increase in the accessibility of a genomic region under the alternative allele has more impact than a decrease, whereas for Enformer the inverse relationship holds. Besides, ChromBPNet scores are sorted according to the fraction of a particular cell type QTLs in the dataset, indicating better alignment of scores to the specific cell type.

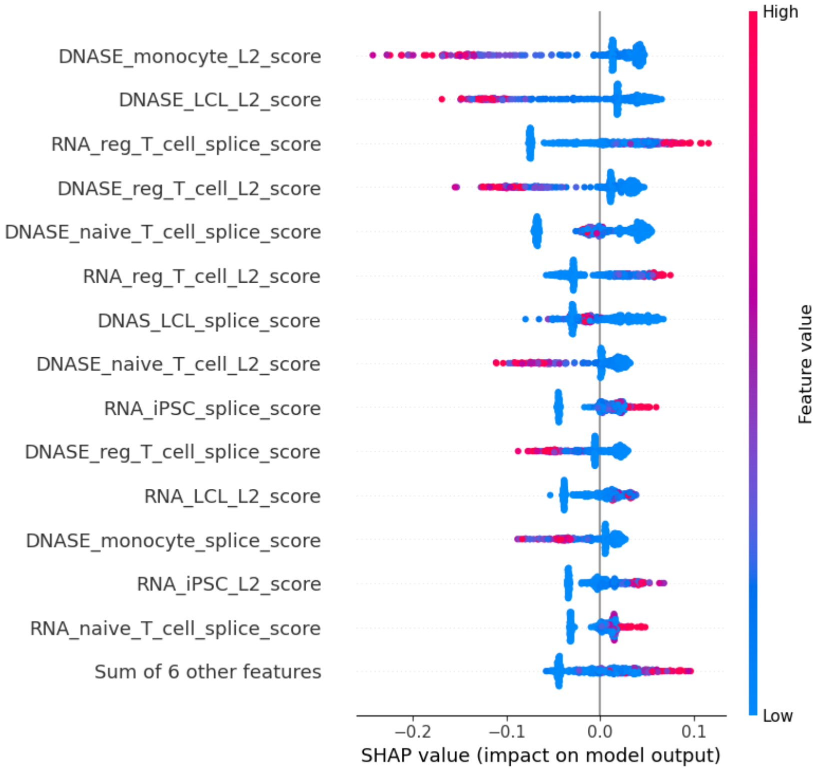


Figure 22. SHAP values plot, displaying a summary of how the top features impact the Random Forest classifier fitted on Borzoi features. Each dot on each row represents one variant from the test set.

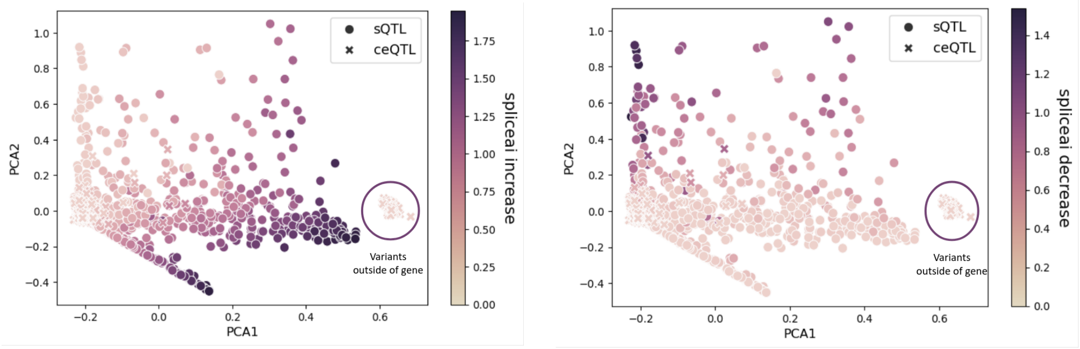


Figure 23. Second and third PCA components derived from the MoA model features in the MoA Dataset. Samples are coloured by the SpliceAI increase feature (left) and SpliceAI decrease feature (right).

Figure 22 shows the SHAP values for genomic features obtained from the Borzoi model. At this point, we would like to remind the reader that the practical difference between the L2 score and the splice score is that the splice score is calculated within the gene boundaries. The SHAP plot analysis shows that higher values of DNASE-related features are associated with predictions aligning more closely with the ceQTL class. Conversely, higher splice scores, derived from RNA coverage tracks, are influential in classifying a variant as an sQTL. The main difference with our model is the fact that for our model, the most influential are splicing predictions, while the Borzoi-based classifier most relies on DNASE tracks prediction scores.

Finally, we performed a PCA analysis of the MoA model features for the MoA dataset and selected eQTL catalogue QTLs (Figures 23, 24 and 25). From Figure 23, we can readily observe a cluster of ceQTLs located outside of gene boundaries and another cluster on the left with the SpliceAI splicing score equal to 0. Overall, it is evident that the PCA dimensions closely align with the SpliceAI predictions.

In the analysis of the eQTL catalogue variants, the picture is not that clear. Notably, some of the Leafcutter variants are clustered with those located outside of gene boundaries, and gene expression QTLs are characterized by high SpliceAI scores, indicating a significant splicing effect.

Overall, the results demonstrate that with the biochemical assays available today, it is evident that specialized, single-task genomic models either match or surpass the performance of large-scale multi-task models, as shown by benchmarking on high-quality datasets. This highlights the effectiveness of tailored approaches in genomic analysis. Sequence-based neural network architectures benefit considerably from integrating dataset-specific features and understanding the limitations inherent in biochemical as-

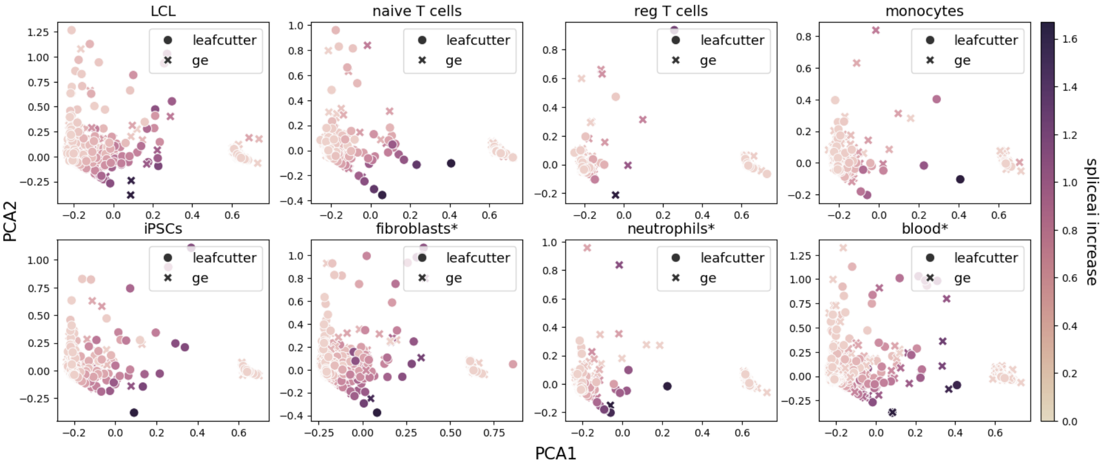


Figure 24. Second and third PCA components derived from the MoA model features in the eQTL catalogue variants set. Samples are coloured by the SpliceAI increase feature.

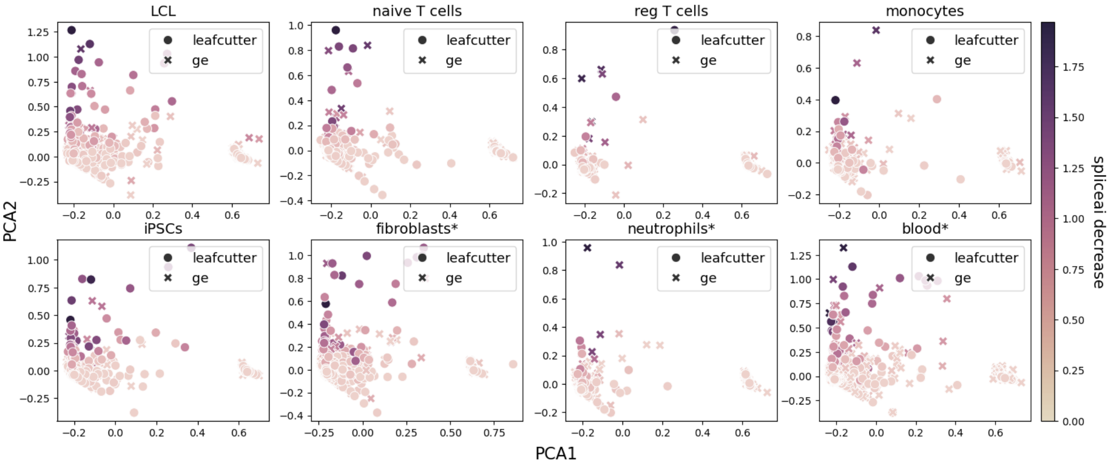


Figure 25. Second and third PCA components derived from the MoA model features in the eQTL catalogue variants set. Samples are coloured by the SpliceAI decrease feature.

says, as shown by the ChromBPNet models. Additionally, modelling multiple regulatory layers concurrently poses significant challenges in the architecture design and training of multi-task models.

The reliance on purely deep learning-based scoring of variants is problematic, often proving to be fickle and unstable. Therefore, the scoring has to be backed up by solid biological or sequence features. The combination of the two types of features allows to distinguish between two very different modes of action of genetic variants (splicing and gene expression change under altered chromatin structure) with almost 90% accuracy, which is not achievable with only classic or only neural features.

Finally, QTL mapping alone as a tool for determining a variant's mode of action is unreliable, frequently resulting in numerous false positives.

6 Discussion

In this study, a significant portion of time was dedicated to the manual labelling of coverage plots for the MoA dataset. Despite these efforts, the number of labelled samples remained too small to effectively train deep learning models. Despite the fact that automated methods have not yet achieved the level of complex decision-making and reasoning that human experts offer, Future efforts could benefit from semi-supervised learning techniques such as pseudo-labelling to mitigate this issue.

In addition to that, we developed the MoA model using biological and deep learning intuition and did not perform any feature engineering based on the MoA dataset, thus demonstrating that the MoA model can successfully tell apart sQTLs and ceQTLs. However, we are well aware that the results are transferable only to QTLs listed in the eQTL catalogue and detected as causal with the fine-mapping and visualization methods used in it. Ideally, an expanded testing set would be necessary to eliminate potential biases and improve model generalization.

Another challenge in genomic research is the complexity and indecipherability of genomic data to humans. As a result, researchers must depend on statistical methods to obtain samples and labels for developing machine learning models. Nonetheless, these methods are susceptible to errors and possess inherent limitations. It is important to remember this fact, especially when ML models trained on the data produced by these methods are later used to refine the results.

One might argue that even though Borzoi scores showed less discriminative power than a hand-crafted set of features, using a single model for predictions is a more convenient and less error-prone approach than juggling a whole collection of models. However, as these models are significantly smaller in size, they can be easily adapted to new cell types or datasets by fine-tuning or training from scratch. Meanwhile, as reported by the Borzoi authors, it took them 25 days and 2 Nvidia A100 GPUs to train a single model.

To evaluate their model on sQTL classification task, the authors of Borzoi constructed a set of sQTLs out of the eQTL catalogue by selecting QTLs detected as txrevise contained event. However, as we showed in this work, QTL mapping is prone to producing a lot of false positives. Therefore, there is a need for more refined high-confidence benchmarking datasets, which is becoming more dire as new, even larger scale and more complex models enter the game. With this work, we made a first step towards that goal.

sQTLs and ceQTLs are, of course, not the only possible molecular mechanisms through which a variant can affect complex phenotypes. Some of them, such as puQTL or apaQTL, can be detected by the quantification methods used in the eQTL catalogue, while some, such as methylation QTL, histone modification QTL, or protein abundance QTL, require different types of assays and methods. Besides, molQTLs can also be shared, as in strong eQTL affecting local splicing or histone modifications affecting gene expression. Finally, some of the effect pathways cannot be explained by any of the most common molecular modalities. Thus, the MoA model serves more like a proof-of-concept method and the set of features used in it is not extensive enough to indicate variants which are not causal or with ambiguous mode of action. So, it is crucial that future studies in this direction continue to expand the set of modes of action detected while still maintaining the possibility of the GWAS variant being assigned to the unknown or ambiguous QTL class.

Lastly, while in this work we adopted an assumption of sQTLs being cell-type agnostic (because of the available datasets), this is not entirely the case. Cellular context is important, and we hope to add it in future work.

6.1 Future work

First of all, we would like to extend the MoA dataset with new molecular traits and cell type specific sQTLs.

Secondly, it would be beneficial to augment the MoA model with the ability to detect variants with indeterminate mode of action. To that end, we can use some anomaly detection techniques or equip the model to express uncertainty in its predictions.

Thirdly, we would like to further explore the cell type specific approach advocated by ChromBPNet. A promising direction is to train ChromBPNet models for more cell types and then map the prediction to a common latent space, where we can explore the similarities and differences in chromatin structure between different cell types, akin to the work by Chen et al. [48].

Finally, to refine our understanding of the associations between genetic variants and molecular traits and build higher-quality benchmarking datasets, we plan to integrate allelic fold change measurements with the already calculated effect sizes and Posterior Inclusion Probabilities (PIPs). This integration can improve the precision in pinpointing functionally significant alleles.

7 Conclusion

Understanding the molecular pathways via which the GWAS variant affects the complex trait can provide useful information about the mechanisms behind various diseases and aid in target prioritization. However, molQTL mapping, which is typically used to assign the variant mode of action, produces numerous false positives and does not work with low-frequency variants. Therefore, in this work, we explored the possibility of using machine learning models to predict the variant's mode of action. We collected the MoA dataset, which includes two classes of molQTLs: splicing QTL and gene expression influenced by chromatin accessibility QTL. In parallel, we compared the performance of two deep learning models, Enformer and ChromBPNet, which represent two opposite approaches to predicting regulatory activity, on a set of fine-mapped chromatin activity QTLs. ChromBPNet proved to be more precise in predicting the caQTLs effect. Finally, we built the MoA model, combining classic genomic features and predictions of single-task deep learning models. The model demonstrated nearly 90% accuracy in distinguishing between the two QTL classes, compared to the 80% accuracy achieved by a classifier based on scores from a single large-scale foundational model. Finally, we scored the QTLs from the eQTL catalogue, detected by either gene expression or Leafcutter methods, with our model. This analysis revealed that while predictions from the MoA model more or less align with gene expression QTLs, most of the Leafcutter QTLs are not classified as sQTLs.

All in all, this thesis presented an original dataset for training and evaluation of the mode of action prediction models and a proof-of-concept MoA model, classifying GWAS variants into two classes: splicing QTLs and gene expression affected by chromatin accessibility QTLs.

8 Acknowledgements

Firstly, I want to express my endless gratitude to the Armed Forces of Ukraine, who kept my family and friends safe, allowing me to focus on this thesis.

Then, I want to say thank you to the members of the computational genomics group for a lot of valuable discussions, which made me realize how little I understand biology and inspired me to try to understand more. Special thanks to Kristiina Kuningas for compiling a dataset on which a part of this thesis is built.

Finally, I want to appreciate the songwriting talents of the Kurgan & Agregat band, whose songs provided me with mental and emotional support during the whole thesis writing process.

References

[1] Jonathan Pritchard. *An Owner's Guide to the Human Genome: an introduction to human population genetics, variation and disease*. Stanford, 2023. URL: https://web.stanford.edu/group/pritchardlab/HGbook.html.

[2] Eric Vallabh Minikel et al. "Refining the impact of genetic evidence on clinical success". In: *Nature* (2024).

[3] Nurlan Kerimov et al. "eQTL Catalogue 2023: New datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs". In: *PLOS Genetics* 19 (2023).

[4] Ralf Tambets et al. "Extensive co-regulation of neighbouring genes complicates the use of eQTLs in target gene prioritisation". In: *bioRxiv* (2024).

[5] John A. Morris et al. "Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens". In: *Science* 380 (2023).

[6] Charles P. Fulco et al. "Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations". In: *Nature genetics* 51 (2019), pp. 1664–1669.

[7] Nurlan Kerimov et al. "A compendium of uniformly processed human gene expression and splicing quantitative trait loci". In: *Nature Genetics* 53 (2021), pp. 1290–1299.

[8] Kaur Alasoo et al. "Genetic effects on promoter usage are highly context-specific and contribute to complex traits". In: *eLife* 8 (Jan. 2019), e41673.

[9] Yang I. Li et al. "Annotation-free quantification of RNA splicing using LeafCutter". In: *Nature genetics* 50 (2017), pp. 151–158.

[10] Benjamin Fair et al. "Cryptic splicing mediates genetic and therapeutic perturbation of human gene expression levels". In: *bioRxiv* (2023).

[11] Hakhamanesh Mostafavi et al. "Systematic differences in discovery of genetic effects on gene expression and complex traits". In: *Nature Genetics* 55 (2023), pp. 1866–1875.

[12] Alvaro N. Barbeira et al. "Exploiting the GTEx resources to decipher the mechanisms at GWAS loci". In: *Genome Biology* 22 (2021).

[13] Ryan Poplin et al. "A universal SNP and small-indel variant caller using deep neural networks". In: *Nature Biotechnology* (2018).

[14] David R. Kelley, Jasper Snoek, and John L. Rinn. “Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks”. In: *Genome Research* 26 (2015), pp. 990–999.

[15] Žiga Avsec et al. “Effective gene expression prediction from sequence by integrating long-range interactions”. In: *Nature Methods* 18 (2021), pp. 1196–1203.

[16] Johannes Linder et al. “Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation”. In: *bioRxiv* (2023).

[17] Kishore Jaganathan et al. “Predicting Splicing from Primary Sequence with Deep Learning”. In: *Cell* 176 (2019), 535–548.e24.

[18] Tony Zeng and Yang I. Li. “Predicting RNA splicing from DNA sequence using Pangolin”. In: *Genome Biology* 23 (2021).

[19] Johannes Linder et al. “Deciphering the impact of genetic variation on human polyadenylation using APARENT2”. In: *Genome Biology* 23 (2022).

[20] Anusri Pampari et al. *Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants.* Version 0.1.1. Jan. 2023. DOI: 10.5281/zenodo.7567627. URL: https://github.com/kundajelab/chrombpnet.

[21] Žiga Avsec et al. “Base-resolution models of transcription factor binding reveal soft motif syntax”. In: *Nature genetics* 53 (2019), pp. 354–366.

[22] Sean Whalen et al. “Navigating the pitfalls of applying machine learning in genomics”. In: *Nature Reviews Genetics* 23 (2021), pp. 169–181.

[23] Sandy L. Klemm, Zohar Shipony, and William James Greenleaf. “Chromatin accessibility and the regulatory epigenome”. In: *Nature Reviews Genetics* 20 (2019), pp. 207–220.

[24] Lingyun Song et al. “Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.” In: *Genome research* 21 10 (2011), pp. 1757–67.

[25] Nurlan Kerimov. “Building a catalogue of molecular quantitative trait loci to interpret complex trait associations”. PhD thesis. University of Tartu, 2023.

[26] W. A. C. Brown and Suzanne Clancy. “Translation: DNA to mRNA to Protein”. In: *Nature Education* 1 (2008).

[27] Suzanne Clancy. “RNA Splicing: Introns, Exons and Spliceosome”. In: *Nature Education* 1 (2008).

[28] Eric T. Wang et al. “Alternative Isoform Regulation in Human Tissue Transcriptomes”. In: *Nature* 456 (2008), pp. 470–476.

[29] Matthias W. Hentze et al. “A brave new world of RNA-binding proteins”. In: *Nature Reviews Molecular Cell Biology* 19 (2018), pp. 327–341.

[30] Gordon L. Hager, James G. McNally, and Tom Misteli. “Transcription dynamics.” In: *Molecular cell* 35 6 (2009), pp. 741–53.

[31] Altuna Akalin. *Computational Genomics with R*. github, 2020. URL: `http://compgenomr.github.io/book/`.

[32] Taras K. Oleksyk et al. “A global reference for human genetic variation”. In: *Nature* 526 (2015), pp. 68–74.

[33] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1 (2021).

[34] François Aguet et al. “Molecular quantitative trait loci”. In: *Nature Reviews Methods Primers* 3 (2023).

[35] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. “From genome-wide associations to candidate causal variants by statistical fine-mapping”. In: *Nature Reviews Genetics* 19 (2018), pp. 491–504.

[36] Gao Wang et al. “A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.5 (2020), pp. 1273–1300.

[37] Masahiro Kanai et al. “Meta-analysis fine-mapping is often miscalibrated at single-variant resolution”. In: *Cell Genomics* 2.12 (2022), pp. 100–210.

[38] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. “A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies”. In: *PLoS Genetics* 5 (2009).

[39] Simone Rubinacci et al. “Efficient phasing and imputation of low-coverage sequencing data using large reference panels”. In: *Nature Genetics* 53 (2020), pp. 120–126.

[40] Frank Rosenblatt. “Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms.” In: *American Journal of Psychology* 76 (1963), p. 705.

[41] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: `http://www.deeplearningbook.org`.

[42] Yann LeCun et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1 (1989), pp. 541–551.

[43] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *North American Chapter of the Association for Computational Linguistics*. 2019.

[44] Ashish Vaswani et al. "Attention is All you Need". In: *Neural Information Processing Systems*. 2017.

[45] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536.

[46] David R. Kelley et al. "Sequential regulatory activity prediction across chromosomes with convolutional neural networks". In: *Genome Research* 28 (2018), pp. 739–750.

[47] Nicholas Bogard et al. "A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation". In: *Cell* 178 (2019), 91–106.e23.

[48] Kathleen M. Chen et al. "A sequence-based global map of regulatory activity for deciphering human genetics". In: *Nature Genetics* 54 (2021), pp. 940–949.

[49] Jasper Janssens et al. "Decoding gene regulation in the fly brain". In: *Nature* 601 (2021), pp. 630–636.

[50] Daniel Quang and Xiaohui S. Xie. "FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data". In: *bioRxiv* (2017).

[51] Jian Zhou and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model". In: *Nature Methods* 12 (2015), pp. 931–934.

[52] 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* 24 (2022).

[53] Alexander Sasse et al. "How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks?" In: *bioRxiv* (2023).

[54] Carl G. de Boer and Jussi Taipale. "Hold out the genome: a roadmap to solving the cis-regulatory code". In: *Nature* 625 (2023), pp. 41–50.

[55] Jennifer A. Doudna and E. Charpentier. "The new frontier of genome engineering with CRISPR-Cas9". In: *Science* 346 (2014).

[56] Haydar A. Frangoul et al. “CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia.” In: *The New England journal of medicine* (2020).

[57] Guillaume Lettre and Daniel E. Bauer. “Fetal haemoglobin in sickle-cell disease: from genetic epidemiology to new therapeutic strategies”. In: *The Lancet* 387 (2016), pp. 2554–2564.

[58] Rory Stark, Marta Grzelak, and James Hadfield. “RNA sequencing: the teenage years”. In: *Nature Reviews Genetics* 20 (2019), pp. 631–656.

[59] Peter J. Park. “ChIP-Seq: Advantages and Challenges of a Maturing Technology”. In: *Nature Reviews Genetics* 10.10 (2009), pp. 669–680.

[60] Andrew J. Bannister and Tony Kouzarides. “Regulation of chromatin by histone modifications”. In: *Cell Research* 21 (2011), pp. 381–395.

[61] Gary Chung Hon et al. “Predictive chromatin signatures in the mammalian genome”. In: *Human molecular genetics* 18 R2 (2009), R195–201.

[62] Zhibin Wang et al. “Combinatorial patterns of histone acetylations and methylations in the human genome”. In: *Nature Genetics* 40 (2008), pp. 897–903.

[63] Feng Yan et al. “From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis”. In: *Genome Biology* 21 (2020).

[64] Alexandre Fort et al. “MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets”. In: *Bioinformatics* 33 (2017), pp. 1895–1897.

[65] Natsuhiko Kumasaka, Andrew J. Knights, and Daniel J. Gaffney. “High resolution genetic mapping of putative causal interactions between regions of open chromatin”. In: *Nature genetics* 51 (2018), pp. 128–137.

[66] Natsue Omi et al. “Efficient and reliable establishment of lymphoblastoid cell lines by Epstein-Barr virus transformation from a limited amount of peripheral blood”. In: *Scientific Reports* 7 (2017).

[67] Lawrence Sie, Susan Loong, and Eng King Tan. “Utility of lymphoblastoid cell lines”. In: *Journal of Neuroscience Research* 87 (2009).

[68] Marianne K. DeGorter et al. “Transcriptomics and chromatin accessibility in multiple African population samples”. In: *bioRxiv* (2023).

[69] Lu Chen et al. “Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells”. In: *Cell* 167 (2016), 1398–1414.e24.

[70] Bruce Alberts et al. *Molecular Biology of the Cell, 7th edition*. W. W. Norton Company, 2022.

[71] Lara Bossini-Castillo et al. “Immune disease variants modulate gene expression in regulatory CD4+ T-cells”. In: *Cell Genomics* 2.4 (2022), p. 100117.

[72] Athanasia D. Panopoulos et al. “iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types”. In: *Stem Cell Reports* 8.4 (2017), pp. 1086–1100.

[73] Tuuli Lappalainen et al. “Transcriptome and genome sequencing uncovers functional variation in humans”. In: *Nature* 501 (2013), pp. 506–511.

[74] Alfonso Buil et al. “Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins”. In: *Nature Genetics* 47 (2014), pp. 88–91.

[75] Benjamin B. Sun et al. “Genomic atlas of the human plasma proteome”. In: *Nature* 558 (2018), pp. 73–79.

[76] Pedro Barbosa et al. “Computational prediction of human deep intronic variation”. In: *GigaScience* 12 (2022).

[77] Lu Chen et al. “Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells”. In: *Cell* 167 (2016), 1398–1414.e24.

[78] Jacob F. Degner et al. “DNaseI sensitivity QTLs are a major determinant of human expression variation”. In: *Nature* 482 (2011), pp. 390–394.

[79] Ankita Arora et al. “The Role of Alternative Polyadenylation in the Regulation of Subcellular RNA Localization”. In: *Frontiers in Genetics* 12 (2022).

[80] Adam Frankish et al. “GENCODE reference annotation for the human and mouse genomes”. In: *Nucleic Acids Research* 47 (2018), pp. D766–D773.

[81] Craig I. Dent et al. “Quantifying splice-site usage: a simple yet powerful approach to analyze splicing”. In: *NAR Genomics and Bioinformatics* 3 (2021).

[82] ENCODEConsortium and Martin Renqiang Min. “An Integrated Encyclopedia of DNA Elements in the Human Genome”. In: *Nature* 489 (2012), pp. 57–74.

[83] Christopher Wilks et al. “recount3: summaries and queries for large-scale RNA-seq expression and splicing”. In: *Genome Biology* 22 (2021).

[84] Zefeng Wang and Christopher B. Burge. “Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.” In: *RNA* 14 5 (2008), pp. 802–13.

[85] Eric L. Van Nostrand et al. “A large-scale binding and functional map of human RNA-binding proteins”. In: *Nature* 583 (2020), pp. 711–719.

[86] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: `https://www.tensorflow.org/`.

[87] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32* (2019). URL: `https://pytorch.org/`.

[88] Paolo Di Tommaso et al. “Nextflow enables reproducible computational workflows”. In: *Nature Biotechnology* 35 (2017), pp. 316–319.

[89] Maya Ghoussaini et al. “Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics”. In: *Nucleic Acids Research* 49 (2020), pp. D1311–D1320.

[90] Natsuhiko Kumasaka, Andrew J. Knights, and Daniel J. Gaffney. “Fine-mapping cellular QTLs with RASQUAL and ATAC-seq”. In: *Nature genetics* 48 (2015), pp. 206–213.

[91] Daniel A. Skelly et al. “Mapping the Effects of Genetic Variation on Chromatin State and Gene Expression Reveals Loci That Control Ground State Pluripotency.” In: *Cell stem cell* (2020).

[92] Dan Liang et al. “Cell-type specific effects of genetic variation on chromatin accessibility during human neuronal differentiation”. In: *Nature neuroscience* 24 (2020), pp. 941–953.

[93] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Dzvenymyra-Marta Yarish**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Predicting the molecular mechanisms of genetic variants,
   *(*title of thesis)

   supervised by Kaur Alasoo.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Dzvenymyra-Marta Yarish
15/05/2024