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

Kristiina Kuningas

Estimating Concordance Between Measured and Predicted Genetic Variant Effects on Chromatin Accessibility

Master's Thesis (15 ECTS)

Supervisor Kaur Alasoo, PhD

Tartu 2023

Estimating Concordance Between Measured and Predicted Genetic Variant Effects on Chromatin Accessibility

Abstract:
Many GWAS studies have identified genetic variants associated with human traits or diseases. However, understanding the underlying molecular mechanisms of those associations has been challenging. Chromatin accessibility is one of those traits that has been associated with a higher risk for a disease. If chromatin is not accessible, then transcription factors cannot bind to it and gene expression or protein synthesis cannot be initiated. This can lead to an altered risk for some diseases. Therefore, it is essential to study quantitative trait loci that affect chromatin accessibility (caQTLs). One of the approaches to find genetic variants is caQTL mapping. It uses open chromatin data and genotype imputation to find associations between genetic variants and chromatin accessibility. Additional fine-mapping distinguishes the potentially causal variants. In addition, deep learning models predicting genetic variants' effects on molecular traits have been integrated into the studies to understand even better the biological mechanisms behind the associations between genetic variants and phenotypic traits. However, the predictive accuracy of these models is still unclear. In this thesis, we created five caQTL datasets for five different cell types based on the fine-mapping results. These datasets were then used to validate the performance of a state-of-the-art deep learning model Enformer in predicting genetic variant effects on chromatin accessibility. Although other studies have evaluated Enformer predictions already, then they have done it from gene expression perspective based on measured effects from RNA-seq data. This thesis, however, compares measured genetic variants' effects on chromatin accessibility from ATAC-seq data to Enformer's predicted effects. It compares both the effect size but also the direction of it. It provides an initial overview of how Enformer performs on chromatin accessibility. Results showed that Enformer performs pretty well on especially the variants for which it predicts stronger effects. In addition, it provided expected results when the cell type of a measured variant was different from the cell type of the predicted variant, meaning it had more opposite effects than it would have with a similar cell type. On the other hand, it also showed very low near-zero effect scores in many cases when the measured effect was higher.

Keywords:
bioinformatics, caQTLs, chromatin accessibility

CERCS:
B110 Bioinformatics, medical informatics, biomathematics, biometrics

Kromatiini avatust mõjutavate geneetiliste variantide mõõdetud ja ennustatud efekti ühildumise hindamine

Lühikokkuvõte:
Genoomiülesed assotsiatsiooniuuringud on tuvastanud mitmeid geneetilise variante, mida on seostatud mõne tunnuse või haigusega. Samas on nende assotsiatsioonide aluseks olevatest molekulaarsetest tunnustest arusaamine keeruline. Kromatiini avatusega seotud kvantitaiivsete tunnuste lookuste (caQTL) uuringud on näidanud, et kromatiini avatus on üks nendest tunnustest, mida on seostatud suurenenud haiguse riskiga. Kui kromatiin pole ligipääsetav, siis ei saa transkriptsioonifaktorid sinna seonduda ning seega ei saa ka geen ekspresseeruda ega proteiini süntees toimuda. See võibki viia suurenenud riskini mõne haiguse osas. Seetõttu on oluline uurida kromatiini avatusega seotud kvantitatiivsete tunnuste lookuseid. Üks lähenemistest, kuidas neid seoseid leida, on caQTL kaardistamine. See meetod kasutab kromatiini avatuse andmeid ning genotüüpide imputeerimist, et leida seoseid QTLide ja kromatiini avatuse vahel. Täppiskaardistamine aitab nende seoste seast leida potentsiaalselt põhjuslikud variandid. Selleks, et veelgi paremini mõista geneetiliste variantide ja fenotüübi tunnuste seoste aluseks olevaid bioloogilisi mehhanisme, on uuringutesse lisatud ka süvaõppemudeleid, mis ennustavad geneetilise variandi mõju molekulaarsetele tunnustele. Küll aga on nendes mudelites kaheldud nende ennustuste täpsuse osas. Selles töös luuakse täppiskaardistamise tulemuse põhjal viis caQTL andmestikku viiele erinevale rakutüübile. Neid andmestikke kasutatakse ka hetkel parima, Enformeri mudeli ennustuste valideerimiseks.Olgugi et ka teised on valideerinud Enformeri mudeli täpsust, on nad teinud seda geeni ekspressiooni vaatepunktist RNA andmetel. Siin töös võrdleme aga täppiskaardistamise mõõdetud geneetiliste variantide mõõdetuid mõju suuruste tulemusi kromatiini avatusele Enformeri mõju suuruste ennustustega kromatiini avatusele. Võrdleme nii mõju suurust kui ka suunda. See töö annab esialgse ülevaate, kuidas Enformer kromatiini andmetel töötab. Tulemused näitasid, et Enformer ennustab päris hästi, eriti kui ta ennustab variandile suuremat mõju. Lisaks sellele olid Enformeri tulemused oodatavad ka siis, kui võrreldi ühe raku mõõdetud efekte sellest erineva raku ennustatud efektidega ehk vastupidiseid efekte oli rohkem kui oleks sarnase rakutüübiga. Teisalt aga ennustas Enformer ka palju nullilähedasi geneetilise variandi efekti mõjusid, kuigi mõõdetud mõju oli suurem.

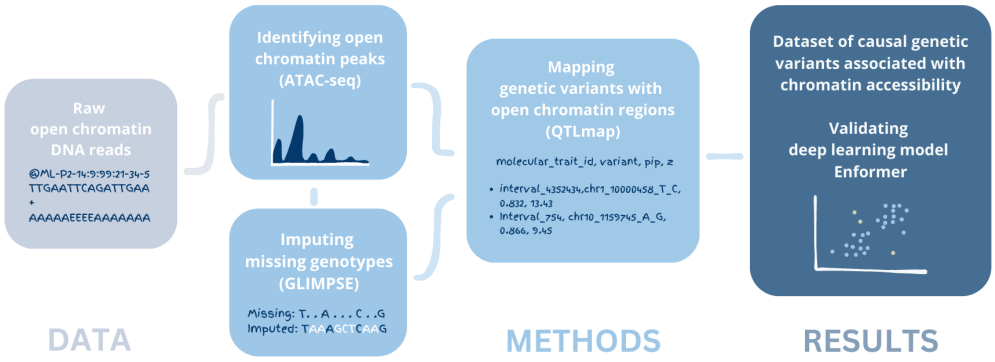
Võtmesõnad:
bioinformaatika, caQTLs, kromatiini avatus

CERCS:
B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Estimating Concordance Between Measured and Predicted Genetic Variant Effect on Chromatin Accessibility

Author: Kristiina Kuningas
Supervisor: Kaur Alasoo, PhD
Data Science (MSc), 2023

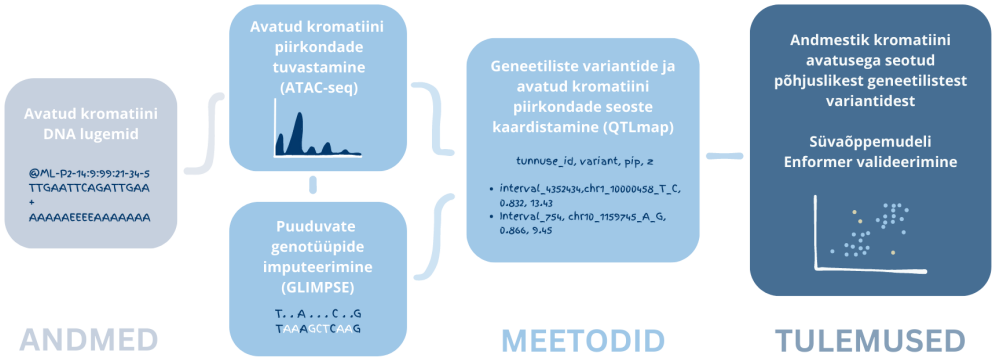
#UniTartuCS



Kromatiini avatust mõjutavate geneetiliste variantide mõõdetud ja ennustatud efekti ühildumise hindamine

Autor: Kristiina Kuningas
Juhendaja: Kaur Alasoo, PhD
Andmeteadus (MSc), 2023

#UniTartuCS



Contents

1 Introduction

Understanding how genotypes become phenotypes is one of the fundamental goals in biology [30]. Genome-wide association studies (GWAS) have found many genetic variants associated with several diseases over the years [80]. However, studying and understanding the underlying molecular processes of these associations is complex as disease-associated genetic variation usually lay in non-coding regions of the DNA [7]. When coding regions (genes) make up approximately 1% of the genome and have been studied thoroughly, then the remaining 99% is yet to be understood [66].

One of the approaches to explain this complex situation, is molecular quantitative trait locus (molQTL) mapping [7]. In molQTL studies, genotypes are mapped with molecular traits such as gene expression, chromatin accessibility or transcription-factor binding [7]. In this thesis, we will concentrate on chromatin accessibility QTLs (caQTLs), which are genetic variants that affect chromatin accessibility. If transcription factors can bind to chromatin, then they make it more accessible, however, it may not always be possible [85]. The state of the chromatin can impact transcription factor binding and it may happen that transcription factors cannot bind to the regions of the chromatin that initiate gene expression for example [85]. Therefore, the gene does not get expressed and protein synthesis does not happen. Because of this, functions of the cells will be affected and this in turn can lead to diseases. For instance, alterations in chromatin accessibility have lead to in vascular smooth muscle cells [64].

Therefore, studying caQTLs that can affect gene expression and potentially increase the risk of developing a disease is essential in the field of genetics. Each additional insight into genetic variants that could be potentially be causal for a molecular trait such as chromatin accessibility is beneficial for understanding the underlying aspects of diseases and possibly providing treatments for those.

Deep learning models that predict the effects of genetic variants on phenotypic traits such as gene expression, chromatin accessibility, etc. have been an addition to other studies to better understand the biological mechanisms behind the associations between genetic variants and phenotypic traits [22,32]. Successful models could theoretically provide us the same results as molQTL studies do, but without having to measure a large amount of individual genetic and molecular profiles [25]. The state-of-the-art model in predicting gene expression is Enformer [25] - a step forward from the previous models because of implementing transformer layers.

However, these models also have their downsides and therefore the accuracy of these models has been confronted in several studies by validating these models from a gene expression perspective [32,33]. However, an overview from a chromatin accessibility perspective would be beneficial to get a better understanding of if this model could be applied in further caQTL research.

Based on the importance of caQTL research, one of the aims of this thesis is to create datasets of likely causal genetic variants affecting chromatin accessibility in five

different cell types. For this, 304 individuals' open chromatin data from five different cell types are used. Cell types are macrophages (naive and stimulated with cytokine interferon-γ stimuli), lymphoblastoid cell lines (LCLs), induced pluripotent stem cells (iPSCs) and endothelial cells (ECs). The open chromatin data from these datasets is then aligned with the reference genome. In addition, missing genotypes are imputed and the QTL-mapping approach is used to map associations between genetic variants and chromatin accessibility. Final caQTL datasets with causal genetic variants are put together from fine-mapping results. As a result of this, five datasets are created with caQTLs in 5 different cell types to provide additional insights into the field.

In addition, two of those datasets are used to validate Enformer, which predicts genetic variants' effects on different phenotypic traits such as chromatin accessibility. Although some studies have already evaluated Enformer from gene expression perspective [32,33], here it will be done from a chromatin accessibility perspective.

As a result of this work, we provide five caQTL datasets of different sizes from different cell types. In addition, an initial insight into how Enformer predicts genetic variants' effects on chromatin accessibility in comparison with measured effects from fine-mapping. Results of this thesis can be used for further studies in the field.

This thesis is divided into six bigger sections. The introduction part gives an overview of the relevance of this thesis' aim. Secondly, an overview of the basics of genetics and the backgrounds of the methods used in further experiments are shared for the reader to comprehend the topic better. In addition, the methodology part describes the datasets used and the underlying details of the workflows used in this work. It also shares details about the technology used in this work. The fourth section shares the results from QTL mapping with fine-mapping and also the results of how well did Enformer predict on caQTLs compared to the measured effects. The discussion section describes and also provides some insight into how the work could be improved. Lastly, a conclusion of the work is provided.

2 Background

In this section, we give an overview of the basics of genetics for a better comprehension of the whole thesis. In addition, the aspects and underlying methods of quantitative trait locus (QTL) analysis and fine-mapping are introduced. What is more, deep learning models in the field are described.

2.1 DNA and genes

Cells are the building blocks of life and all living organisms are made of cells [1]. Human bodies are composed of eukaryotic cells and almost each of those contains genetic material called deoxyribonucleic acid (DNA), which is responsible for the genetic information inheritance of the organism [1]. DNA is a double-stranded helix [2] consisting of two polynucleotide chains, which are composed of four types of nucleotides, adenine (A), cytosine (C), guanine (G) and thymine (T) [2]. The two strands are held together by hydrogen bonds that form between the nucleotides on each of the strands, creating base pairs in the process [2]. Adenine is paired with thymine and guanine with cytosine [2].

Genes are segments of the DNA sequence that encode for proteins [5,1]. Genes are critical for human functioning by affecting cell functions and thereby influencing the traits of individuals [1]. For the proteins to formulate, the gene has to be expressed via gene expression [70].

While genes are the coding regions of DNA that give instructions for gene expression, the more significant portion of the genome is DNA regions that are non-coding. The most recent studies say that around 20000-250000 protein-coding genes in the human genome make up 1% of the whole genome [66] leaving the remaining 99% non-coding. Although the focus in research has been on coding parts of the DNA and non-coding regions have been at some point defined as irrelevant parts of the DNA [67,68], then with the completion of the human genome, they are now looked at as having a very important regulatory role in organisms [69]. Non-coding elements that can regulate gene expression are for example promoters, enhancers, insulators and silencers [69]. These are also called *cis*-regulatory elements (CRE) [10,11,69].

2.1.1 Gene expression

Gene expression encodes the gene into proteins and consists of two main steps, transcription and translation [1]. In the initial stage, transcription, DNA gets transcribed into messenger ribonucleic acid (mRNA) [1].

Transcription starts when an enzyme called RNA polymerase binds to a promoter region near the gene, signals the DNA to unwind the double-helix, and starts transcribing one chain of the DNA (template strand) to pre-messenger RNA (pre-mRNA) [5].

However, for this enzyme to bind and the gene expression to start, proteins called transcription factors have to bind to the promoter region first [1]. The promoter region is just a sequence of DNA upstream of a gene at the end of the transcription start site (TSS) [71] where proteins, such as transcription factors or RNA polymerase itself, can bind to initiate the transcription of that gene [70]. Promoter regions also dictate which strand of the gene will be transcribed [70]. The pre-mRNA is formed when the RNA polymerase adds the RNA nucleotide that matches the DNA nucleotide in the template strand to the pre-mRNA strand [70]. The transcription process ends when termination happens once the enzyme transcribes a DNA sequence called the terminator [70].

In the translation part of the gene expression, mRNA is decoded into protein molecules with the help of another RNA-protein structure called a ribosome [5]. The mRNA is read by the ribosome as the mRNA passes through it [5]. The mRNA acts as a template for forming a protein as it has the genetic code in the form of codons, triplets of nucleotides, which code for a particular amino acid [5]. These codons are then decoded into chains of amino acids which eventually form the proteins [5].

2.2 Chromatin

The whole DNA of an organism, the genome, has to fit into the nucleus of the cell. As the stretched-out genome is approximately two meters long because of the sizes of the nucleotides, the DNA has to be very tightly packaged to fit [72]. To address this issue, DNA is packed into 23 pairs of chromosomes in each of the cells [71], where one pair is inherited from the mother and the other from the individual's father. Chromosomes are made up of a nucleoprotein called chromatin [6].

DNA sequences of the length of 146 base pairs are wrapped around octamers of proteins called histones, forming nucleosome core particles in the process [6]. Repeating nucleosome units make up a fiber called chromatin and this fiber eventually condenses into a chromosome [6,72]. The whole process of how DNA is packed into a chromosome is visually described in Figure 1.

2.2.1 Chromatin accessibility

Chromatin accessibility plays a critical role in impacting cell functions. It is dependent on the physical access of enzymes targeting the DNA segments [85]. If nucleosomes in the chromatin are too densely arranged (closed chromatin), the regulatory proteins cannot attach to it and relevant processes such as gene expression, cannot begin [85]. This affects the functioning of the individual. However, when nucleosomes get depleted, typically at promoters and enhancers (open chromatin), the transcription factors can bind to the DNA and transcription is able to begin.

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-

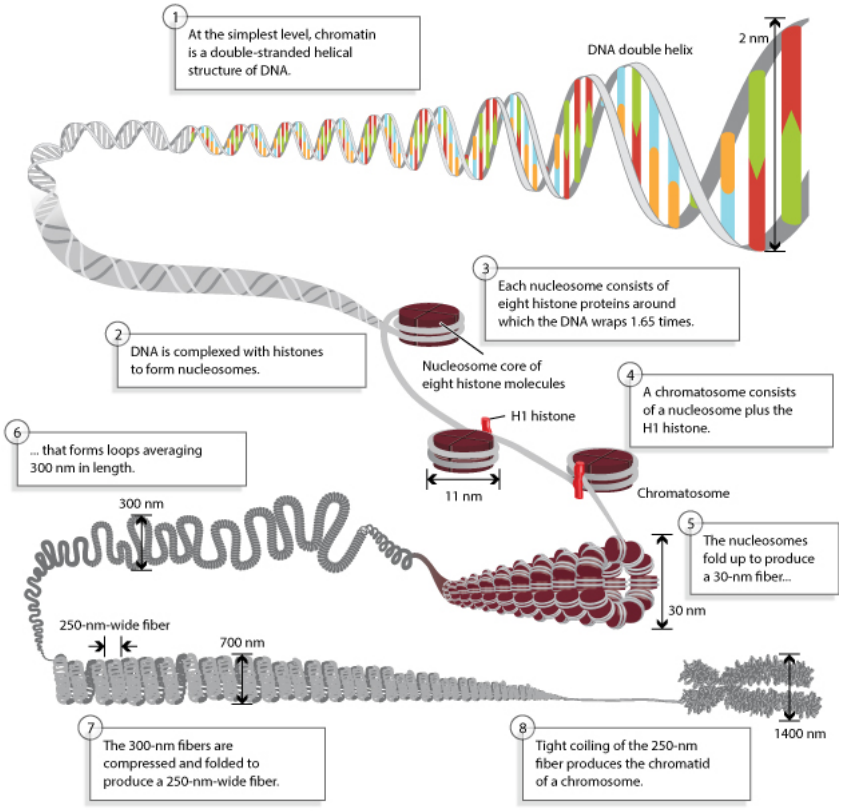


Figure 1. An overview of how DNA is packed into the chromosome [6]. DNA is wrapped around the histones,

wide chromatin accessibility profiling methods available. In 2013, Buenrostro et al. [41] came out with an assay for transposase-accessible chromatin using sequencing (ATAC-seq), which identifies regions of open chromatin. In the initial step of this method, Tn5 transposase with sequencing adaptors, which are loaded into the Tn5, will bind to accessible DNA and cut out accessible segments [41] creating the ATAC-seq library (see Figure 2). Those sequencing-library fragments are then amplified and sequenced [41] and can be used for later analysis. Less accessible chromatin makes this kind of transposition less probable [41] and therefore the transposase can bind only to open chromatin.

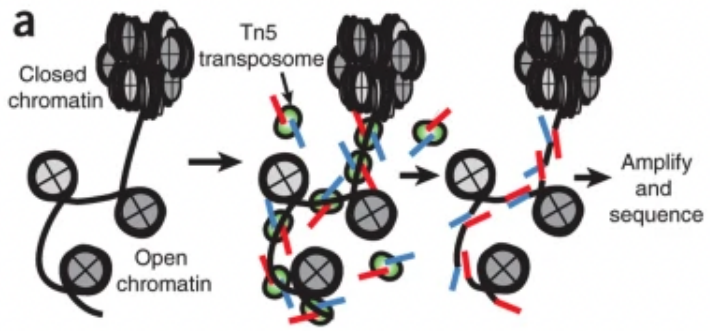


Figure 2. The process of how open chromatin sequences are produced [41]. Tn5 transposase marked with green and loaded with sequencing adaptors marked with blue and red binds to the open chromatin regions of DNA between the grey nucleosomes. During this, accessible DNA segments are cut out and they are amplified and sequenced.

Besides ATAC-seq there are other methods that provide information on chromatin accessibility or transcription-factor binding, such as DNase-seq [42], MNase-seq [43] and FAIRE-seq [44]. However, in comparison with ATAC-seq, those methods require many multiple orders of magnitude more input cells [41]. For example, if ATAC-seq requires 500-50000 cells, then MNase-seq requires 1-10 million and DNase 50 million cells [46]. Needing this many cells excludes the possibility of profiling rare but important cellular sub-types in genome-wide chromatin analyses [41]. In addition, they take a lot more time than the ATAC-seq process. Therefore, after ATAC-seq arrived in 2013, it has become a popular epigenomic analysis method to use based on the fact that the amount of ATAC-seq datasets as well as publications based on those has increased exponentially over the last decade [47]. It has been widely used to study chromatin accessibility or to identify regulatory regions of the genome [47].

Studies have found that changes in chromatin structure, in particular its accessibility, can lead to genetic disorders that eventually can cause diseases. For instance, in 2021 Wang et al. [40] found that the alterations in chromatin accessibility can influence the plasticity of atherosclerotic vascular smooth muscle cells (SMCs) which can negatively impact their response to inflammation-induced stress. They used ATAC-seq data to conduct the study and reach conclusions [40].

Regulatory proteins, such as transcription factors, bind to non-coding regions, where the regulatory elements are situated, to usually initiate gene expression [46, 41]. However, if the DNA is so densely packed that it is unreachable for the regulatory proteins, then genes can not be expressed and protein synthesis can not happen [85]. That can lead to malfunctioning of the cells and eventually the whole organism, which in turn can lead to diseases. Therefore it is essential to study chromatin structure and its accessibility.

2.3 Genetic variants and alleles

It is known that the genetic difference between individuals is only around 0.1% [3]. Genetic variants are different versions of DNA that lead to different genotypes and eventually different phenotypes [1]. Genotype represents the specific combination of alleles in each gene which is inherited from an individual's parents, one allele from the mother and the other from the father [1]. The genotype determines eventually the phenotypic traits of the individual [1]. Phenotypic traits can be physical characteristics but also physiological traits or disease susceptibility [1]. Therefore, this small 0.1% difference defines why individuals look different from each other or why they are more prone to some diseases than others.

Genetic variants are mainly caused by mutations or recombination of genetic material [83]. When there has been a change in the DNA sequence, then it is called a mutation [83]. Recombination, on the other hand, occurs when the genetic material of maternal and paternal DNA is shuffled and rearranged before the cell division resulting in new combinations of genetic variants in the child's germ cells [83].

One common type of genetic variant involving a change in the DNA sequence is single nucleotide polymorphism (SNP), where the mutation happens when one single nucleotide of the genome is changed [1,81]. For example, if cytosine is replaced with thymine. They can be located both in non-coding and coding regions of the genome [81]. It is possible that they do not have any significant impact on an individual's health, however, they can also play a part in influencing a trait [39] or in disease development [10]. Figure 4 shows how a trait can be influenced by a SNP. In addition to SNPs there are also another type of mutations called insertion-deletion polymorphisms (indels) where nucleotides in the sequence are whether inserted or deleted [73].

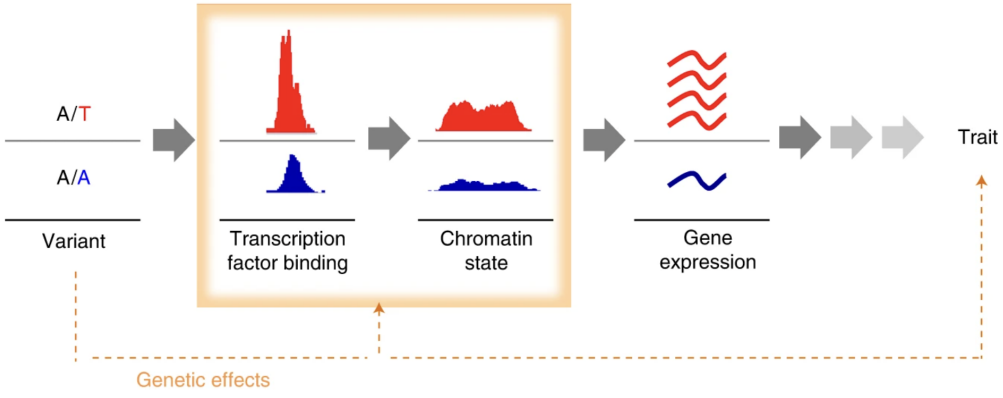


Figure 3. An illustrative example of how genetic effects can affect a trait [39]. Firstly, a SNP, where adenine is replaced with thymine, increases the transcription-factor binding which in turn affects the chromatin state. Chromatin state, however, influences the level of gene expression which leads to affecting the trait.

Quantitative trait locus (QTL) is a genome region directly associated with a quantitative trait or phenotype [7,74]. They are usually whether SNPs or indels. Quantitative traits are measurable phenotypic traits. QTLs associated with molecular traits are called molQTLs [7]. Those molecular traits can be for instance chromatin accessibility or gene expression and QTLs associated with those traits are termed caQTLs and eQTLs accordingly [7].

If SNPs are located in the regulatory regions of genes or directly in genes themselves, then they can directly affect the gene's function and therefore play a noticeable role in the development of the disease [10]. As an example of SNPs, an increased risk of developing breast cancer has been associated with SNPs in BRCA genes [64].

2.4 QTL association mapping

Genome-wide association studies (GWAS) have found that SNPs have been associated with risks for several diseases, however, it is also important to know the molecular processes that cause this kind of effect of these variants [80,7] and molQTL analysis can be used for this. QTL association mapping is a statistical method that links genotypic and phenotypic data to describe the underlying genetic mechanisms that contribute to differences in traits and eventually diseases [12]. It finds the chromosomal regions, such as QTLs, that significantly affect the quantitative traits [7]. From this thesis perspective, QTL mapping helps to determine what are the potential causal caQTLs, which affect chromatin accessibility, which in turn influences transcription factor binding and eventually could result in a higher risk for a disease.

Genetic variants from the QTL analysis with low p-values indicate statistically significant associations with the trait under observation. It could be expected that the variant with the most significant association (lead variant) [16] would be the causal variant, although it may not be the case [9,15,17]. Statistical methods alone cannot determine variant causality. For example, there could be several variants that are in linkage disequilibrium (LD) with some other variant [3,12,15,20]. There are many studies that contain several pairs of genetic variants with correlations that equal even 1 [15]. Therefore, it is hard to determine the independent potential causal variants, because variants could be highly correlated with one another.

To address this, a further fine-mapping step is necessary to pick smaller sets of variants that likely contain independent causal signals amongst the significantly associated variants [7,17]. There exist several different fine-mapping strategies, such as heuristic and penalized regression approaches as well as Bayesian methods [16]. Bayesian methods have been designed directly for fine-mapping and therefore they have advantages over other approaches [16]. Some Bayesian approaches are based on Bayesian variable selection in regression (BVSR) [15]. BVSR can address the uncertainty of which variables to select despite the high correlation of variables [15].

One of the models based on BVSR is the "sum of single effects" model SuSiE [15].

SuSiE produces 'credible sets' of variables that have a certain probability of "containing at least one variable with a non-zero regression coefficient" [15]. Variables in this thesis refer to genetic variants. Each of the variants is given a posterior inclusion probability (PIP) showing how strong the evidence is that this variable is an effect variable or in this case a causal variant [3,15]. The higher the PIP value the more probable it is that this SNP is potentially causative for a trait [15]. Credible sets are formed when the cumulative sum of the PIPs reaches a certain threshold probability, for example 90% or 95% [15,17].

An illustrative example of how credible sets, statistically significant associations and causal variants are related is in Figure 4. The lead variant with the lowest p-value is represented with the yellow and we can see several variants being in high LD with that variant. Despite being the lead variant, it is clear that the variant with the highest PIP value is not the lead variant itself, but rather a variant that is in high LD with it. The credible set of possible causal variants is illustrated with the grey area. This illustrates that the most statistically significant variants might not be the causal ones but in high correlation with them.

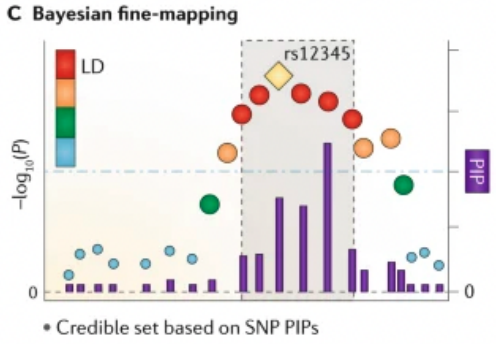


Figure 4. Illustration of Bayesian fine-mapping results, where yellow rhomb represents the lead variant rs12345 and red dots represent the genetic variants in high LD with the lead variant [16]. It can be seen that the actual causal variant with the highest pip-value is not the lead variant with lowest p-value. The area with grey background illustrates one credible set.

2.5 Genotype imputation

Completeness of the genotypic data in molQTL studies is essential for fine-mapping quality because incomplete genotypic data of the potential causal variants could affect fine-mapping results in a way that leads to artefactual associations [7]. Genotype imputation is the process that helps to infer the unobserved genotypes [78]. It is essential for fine-mapping [78] as it expands the number of SNPs for the association mapping and

also increases the possibility of detecting true associations [8].

In the process of imputation, samples are compared against the haplotypes from a big reference panel that has the whole genomes of a bigger population [8]. Haplotypes are combinations of alleles that tend to be inherited together. Samples with missing genotypes are compared against several haplotypes and combinations of those are then imputed into the sample [8].

The reference allele is the allele that appears in the reference genome and is "typical" for the bigger population. The alternate allele, on the other hand, is any allele that is not the reference allele. The alleles could be of the length of one nucleotide, but also longer nucleotide sequences, depending on whether the imputation was based on only the change of a nucleotide or the insertion or deletion of many.

The availability of next-generation sequencing (NGS) has allowed for whole-genome sequencing data for GWAS studies. Genome-wide sequencing (GWS) data can be used for genotype imputation [78]. Compared to the SNP chip method on genotyping arrays [78], where only common variants are included, GWS allows using regions of the genome that are unknown. Therefore, in addition to the common variation, it can capture variations that are rare across the whole genome [78]. However, it is more costly than using SNP arrays [63]. One cost-effective alternative solution for GWS is low-coverage whole-genome sequencing [63]. It provides as much common variation, but more rare variation than standard SNP array platforms [63].

GLIMPSE by Rubinacci et al. [63] is one of the imputation methods designed to impute low-coverage sequences from reference panels using a combination of SNP enrichment and imputation methods. GLIMPSE stands for "genotype likelihoods imputation and phasing method" [63]. It has also proven itself to outperform other methods with its lower computational cost and accuracy [63].

GLIMPSE also incorporates quality control measures to filter out low-quality variants. Quality control in genotype calls is as essential as the completeness of the data [7]. In this thesis, minor allele frequency (MAF) filter is used to filter out rare variants with low allele frequencies because of being less informative for fine-mapping purposes. Another filter, the imputation information (INFO) filter reported by another imputation method IMPUTE2 [8], is used to pick out values with higher imputation accuracy. INFO filter values near 1 indicate a SNP being imputed with high accuracy [65].

One of the studies that has done genotype imputation to find chromatin QTLs (cQTLs) is by Baca et al. [39] where they did it on chromatin immunoprecipitation (ChIP-seq) data. For this thesis, however, genotype imputation is done on ATAC-seq data to find caQTLs. Peep Kolberg has developed a genotype imputation workflow based on GLIMPSE in his master's thesis about eQTLs analysis on single-cell RNA-seq data [57]. He demonstrated that GLIMPSE can also be used to accurately impute genotypes directly from ATAC-seq data [57]. Kolberg's workflow was used in this work's imputation step.

2.6 Deep learning

To gain an even better insight into the complex situation of explaining what are the underlying factors of non-coding genetic variants associated with phenotypic traits, machine learning models have been integrated into the studies [22,32]. Today, deep learning models that are mainly built on convolutional neural networks (CNN) are the state-of-the-art approaches for predicting traits such as gene expression, chromatin accessibility or transcription-factor binding from DNA sequences to explain better the biological mechanisms behind these associations between genetic variants and traits that eventually can result in diseases [32,25]. Successful models should help us to achieve the same results as molQTL studies do, only without having to measure a large amount of individual genetic profiles [25]. Hence, well-performing machine learning models could be an important source for deciphering the *cis*-regulatory code [24] and could be possibly used in designing personalised diagnoses or treatments [33].

These models are usually built on CNNs that take DNA sequences as input and predict functional properties as outputs [32]. One of the shortcomings of many of those models is that they can only make use of DNA elements not more than 20 kb away from the transcription start site (TSS) [25]. However, some regulatory elements such as enhancers to where transcription factors bind to can activate transcriptional activity from much further away. Therefore, these models are "blinded" for elements outside of their receptive field and could not get use of them in making predictions. This is the limitation of using only convolutions in the models and this limitation could be quite critical if researchers want to achieve a wider understanding in the field.

In 2021, Avsec et al. [25] addressed these shortcomings with a model called Enformer, which is also the state-of-the-art model for predicting gene expression as of now [33]. The name comes from a combination of enhancer and transformer [25]. When previous models could reach elements up to 20 kb away, then Enformer expands the receptive field to up to 100 kb [25,33]. The main difference between the previous models is that in addition to the convolutional layers, Enformer also has attention layers, which are characteristic to transformers (see Figure 7). That is why transformers are able to read long texts and therefore also longer DNA sequences. This has led to an increase in prediction accuracy compared to the previous best-performing models like Basenji2 [28] or ExPecto [27].

Enformer takes in 200 kb long DNA sequences of both human and mouse reference genomes [25]. These sequences are then fed to the convolutional layers which prepare the data for the transformer by summarizing the information and compressing it [25]. Pooling layers help to reduce the input sequence dimension to 1536 bp which makes each sequence position vector 128 bp long [25]. As a next step, transformer blocks need to detect long-range interactions information between DNA regions [25]. After transformer layers, two output heads of human and mouse predict genomic tracks that show phenotypic traits for sequential genome positions [25]. There are four tracks predicted by

Enformer: cap analysis gene expression (CAGE), histone modifications (ChIP histone), transcription factor binding (ChIP TF) and DNA accessibility (DNase/ATAC) [25]. In the context of this work, the DNase/ATAC tracks will be used to find how Enformer predicts associations between DNA sequences and open chromatin peaks. The whole process of how input sequences evolve into predictions for the tracks could be seen in Figure 5.

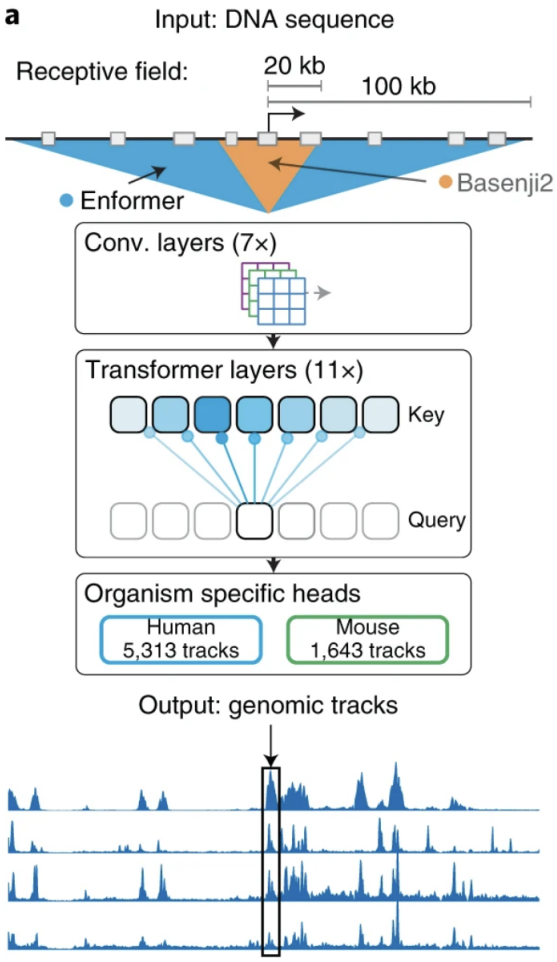


Figure 5. Overview of the Enformer architecture [25]. The initial input for Enformer are the DNA sequences. Secondly, this data is fed to the convolutional layers which in turn are input for the following transformer layers. After transformer layers, two output heads predict four different tracks. These tracks are visualised in the bottom of the Figure.

Despite deep learning models being seen as an important tool to interpret the genetic variations in genomes [32], they have also been confronted in terms of their prediction accuracy [33,22]. Machine learning models usually expect the input training data to be as independent and identically distributed as possible to generalize to new data [22,35],

however, the evolutional concept of genetic data does not allow that. DNA sequences are far away from being unrelated as 99.9% of individuals' genomes are the same [84] and in addition approximately 70% of the genome is repetitive [36]. Therefore genome-trained models can lead to false causal signals resulting in unreliable predictions of genetic variants and regulatory elements' functions.

Enformer is also one of the models which is trained on reference genome sequences [25] and therefore can also not represent the diversity between certain populations or individuals [37]. Thus, it is important to critically evaluate the accuracy of these models. Although Avsec et al. [25] have provided evidence of Enformer's accuracy on gene expression, they did it on data that is highly enriched for cancer cell lines and therefore does not give a full picture of its performance. Karollus et al. [33] and Sasse et al. [32] evaluated Enformer from the gene expression perspective by comparing Enformer predictions on the CAGE track against their observed genetic variant effects from gene expression data. CAGE track was used because itSingle was the most relevant track in terms of gene expression.

Although previous studies have already investigated the quality of Enformer, they have done it from a gene expression perspective. Thus, it would be beneficial to do it from a DNA accessibility perspective as well. Validating Enformer performance in predicting how genetic variants influence chromatin accessibility is a necessary procedure to get a better understanding of if this model could be applied in caQTL research. Just like Sasse et al. [32] and Karollus et al. [33] did for gene expression tracks by comparing Enformer predictions with measured eQTLs gene expression values, predictions on chromatin accessibility tracks could be compared against measured associations from accessible chromatin data.

3 Methodology

In this methodology section we firstly give an overview of the datasets that were used in this thesis' experiments. Furthermore, all of the different workflows used in this thesis are described.

3.1 Data

In 2018, Alasoo et al. [49] identified genetic variants that were associated both with chromatin accessibility and gene expression in human macrophages. They used human induced pluripotent stem cell (IPS) lines from 123 male and female donors [49]. These cells were then used to differentiate macrophages, which are mononuclear cells that protect the organism by being able to participate in phagocytosis [48] and therefore have a significant defensive role in our immune system [48]. For the purpose of Alasoo et al.'s work, they split the macrophages into four subsets by stimulating three of those with different stimuli such as cytokine interferon-γ (IFNg), *Salmonella enterica serovar Typhimurium (Salmonella)* (SL1344), IFNg stimulation followed by Salmonella infection (IFNg+SL1344) [49]. The fourth one remained unaffected, naive [49]. Out of these four stimulated subsets, naive and IFNg ATAC-seq data of accordingly 42 and 41 donors were used in this work. The other two were left out because of having too few samples for this thesis' purpose. This data already existed in the HPC and was available to use for this thesis.

Secondly, in 2018, Kumasaka et al. [51] investigated potential causal interactions between open chromatin regions using chromosome conformation mapping technology and genetic data analysis. They showed that open chromatin regions can communicate with each other and affect gene expression, which in turn can affect the development and onset of diseases [51]. Kumasaka et al. [51] used lymphoblastoid cell lines (LCLs) which are formed when B lymphocytes are infected by one of the most common herpesvirus types, Epstein-Barr virus (EBV) [52]. ATAC-seq data from 100 individuals from the 1000 Genomes Project[1]was applied in their work [50]. 24 samples out of the total 100 were already prepared in their previous research in 2016. In this paper, all of the 100 individuals' ATAC-seq data was used. The data is available in the European Nucleotide Archive[2]

In 2020, Stolze et al. [50] tried to identify regulatory elements in disease-relevant cell types to detect causal variants that would contribute to the development of complex diseases. For this, they used human aortic endothelial cells (ECs) from both sexes and three major ancestries [50]. ECs regulate vascular tone and also create an anti-thrombotic surface to cover the inner walls of arteries, veins and capillaries [50]. They

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

[2]https://www.ebi.ac.uk/ena/browser/home

also participate in the development of complex diseases, for example atherosclerosis [50]. Stolze et al. [50] found thousands of eQTLS unique to endothelial cells that had not been detected in previous studies [50]. For their experiments, they treated the cells with human recombinant IL-1B protein or with no protein at all [50]. This thesis used the ATAC-seq data of the untreated cells of 44 donors.

Lastly, human induced pluripotent stem cells (hiPSCs) were added to the work. The ATAC-seq data for those was taken from iPSC Collection for Omic Research (iPSCORE)[3], which purpose is to provide data for use in studying the impact of genetic variation on molecular and physiological phenotypes. One of the studies that used the ATAC-seq data from those cells is by Greenwald et al [53]. ATAC-seq data of 77 donors of iPSC cells were used in this work.

Altogether this work included data from 304 different donors of four different cell types: macrophages (naive and IFNg), lymphoblastoid cells, endothelial cells and pluripotent stem cells. In this work, the datasets are referred to as Alasoo's naive and IFNg, Stolze's, iPSCORE's and Kumasaka's datasets. Table 1 shows the datasets, corresponding sample sizes and cell types that were used in this work.

Table 1. Overview of the datasets.

Dataset	**Donors**	**Cell type**
Alasoo (IFNg)	41	macrophage_IFNg
Alasoo (naive)	42	macrophage_naive
iPSCORE	77	iPSC
Kumasaka	100	LCL
Stolze	44	EC

3.2 Aligning ATAC-seq reads

The initial step was to identify the areas of accessible chromatin (open chromatin peaks) in the human reference genome . A human reference genome is a genome that represents not only an individual's genome but a genome of a collection of many [76]. The reference genome used in this thesis was GRCh38 [75], the latest one available for human. Raw FASTQ files contained open chromatin region DNA sequences (ATAC-seq data) from various samples described in the previous section. FASTQ is a file format for storing genome sequencing data [54]. Those regions from FASTQ files were then aligned against the GRCh38 genome to determine the regions of that genome where samples had the highest read counts (open chromatin peaks).

This process was done with the help of nf-core Nextflow ATAC-seq pipeline[4] which purpose is to analyse ATAC-seq data. Their ATAC-seq pipeline reads in the raw FASTQ

[3]https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study$_id = phs001325.v3.p1$

input files, aligns the reads, does peak calling and also performs quality control [55]. Within the pipeline, it uses several other common bioinformatics tools, such as Picard, BWA, BEDtools, BAMtools, MACS2 and featureCounts [55]. The pipeline outputted binary alignment map (BAM) files which contained the aligned sequences [55]. After peak-calling, featureCounts was used to create a count matrix, where rows marked the peaks, columns the different samples and the values represented the read counts for that specific sample in a specific peak region [55]. Both of these outputs, BAM files and count matrix are relevant for the following stages of this work.

3.3 Genotype imputation

Genotype imputation was done with the use of a workflow[5], which is based on GLIMPSE [63]. The workflow was developed by Peep Kolberg as part of his master's thesis [57]. The workflow takes in the sequencing read BAM files from the previous step and outputs a variant call format (VCF) file, which contains the information about the variant, the reference and alternate allele and the quality scores of the imputation for each of the samples [57]. For quality control, we used MAF and INFO filters to filter out low-quality genotypes. The MAF filter we used was based on the minor allele count (MAC) filter. We determined a MAC threshold of 6, meaning that more than 6 individuals had to have alternative allele to avoid false-positive fine-mapping results. Therefore also the threshold of the MAF filter varied based on the dataset size. Values varied from 0.01 to 0.07 and the more samples the dataset had, the smaller the MAF used was. INFO filter was 0.4 for all the datasets. Genotype imputation with GLIMPSE was done on Kumasaka's, Stolze's and iPSCORE's datasets. Genotypes for Alasoo's datasets were already available before.

3.4 Relatedness

For two of the datasets, Kumasaka and iPSCORE, additional genotype quality control was done to find out the genetic relatedness between the samples. For the purpose of QTL association mapping, it is necessary, that the input samples would be unrelated [7,12]. The analysis was done with plinkQC[6], which is an R package for genotype quality control. Relatedness matrices were created to see which samples were related and how strong the relatedness was.

The iPSCORE dataset by default contained donors who were from the same families and therefore related. Therefore, it was necessary to keep only one of the samples out of those who were related. As a result of this 57 samples from 134 were eliminated from the

[4]https://nf-co.re/atacseq

[5]https://github.com/peepkolberg/glimpse

[6]https://zenodo.org/record/3934294#.ZFp01uxBz0o

initial iPSCORE dataset and 77 remained. For Kumasaka data, it was just a precaution to check for relatedness as in the beginning more samples than in the Kumasaka et al. [51] paper were available for this work. Therefore, a hypothesis, that some of them could be related, had to be tested. After seeing the results from the matrix, it was clear that the hypothesis became true. In Figure 6 it is clearly seen that 12 of the samples are clearly related to another sample. These 12 samples were then removed from the experiments presented in this thesis.

To conduct the Plink analysis and create the relatedness matrix to filter out related samples, VCF file from the genotype imputation step was needed. Therefore, after cleaning the data from related individuals, ATAC-seq and genotype imputation workflows had to be run again to produce the most correct results for the following QTL analysis.

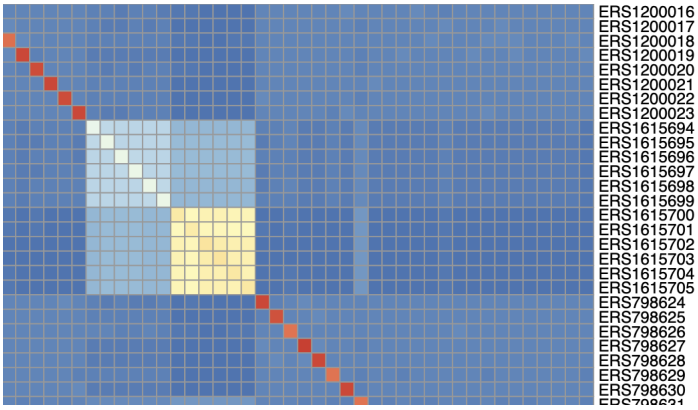


Figure 6. A snippet of the relatedness matrix for Kumasaka dataset samples. Red squares on the diagonal represent each sample's correlation with itself and therefore red represents the highest correlation. Dark blue squares show no or very low correlation. Lighter blue and yellow squares show some relatedness between samples. We can see that there are 12 samples that are related to some other sample, which cannot be seen on this graph.

3.5 QTL mapping

The QTL association mapping was done with an eQTL-Catalogue/qtlmap bioinformatics pipeline[7], which is based on QTLtools[8]. The following fine-mapping analysis is based on the SuSiE model [15] and uses SusieR[9]to apply it. It takes in four types of inputs: trait matrix, trait metadata, sample metadata and genotype VCF.

The trait matrix in this thesis was the normalised version of the featureCounts peak

[9]https://github.com/eQTL-Catalogue/qtlmap
[9]https://qtltools.github.io/qtltools/
[9]https://stephenslab.github.io/susieR/index.html

count matrix from the Nextflow ATAC-seq workflow output. To normalise it, for each sample in the matrix, each of the read counts was divided by the sum of the read counts for that sample. Then all of the data in the matrix was multiplied by a million to get counts per million (CPM). Subsequently, rank-based inverse normal transformation (INT) was applied to the matrix [58]. As a result of these steps, the rather unbalanced data was transformed into a normal distribution.

Trait metadata contains information about the phenotypes, the chromatin accessibility peak regions. It includes information such as the position of the peak, the chromosome on which the peak is situated, the start and end positions of it and also the length of the peak. Sample metadata, on the other hand, is a file that contains all relevant metadata for the samples. For instance, there is info about sample id, sex, cell type, condition and read length. The genotype VCF is the VCF file from the genotype imputation step.

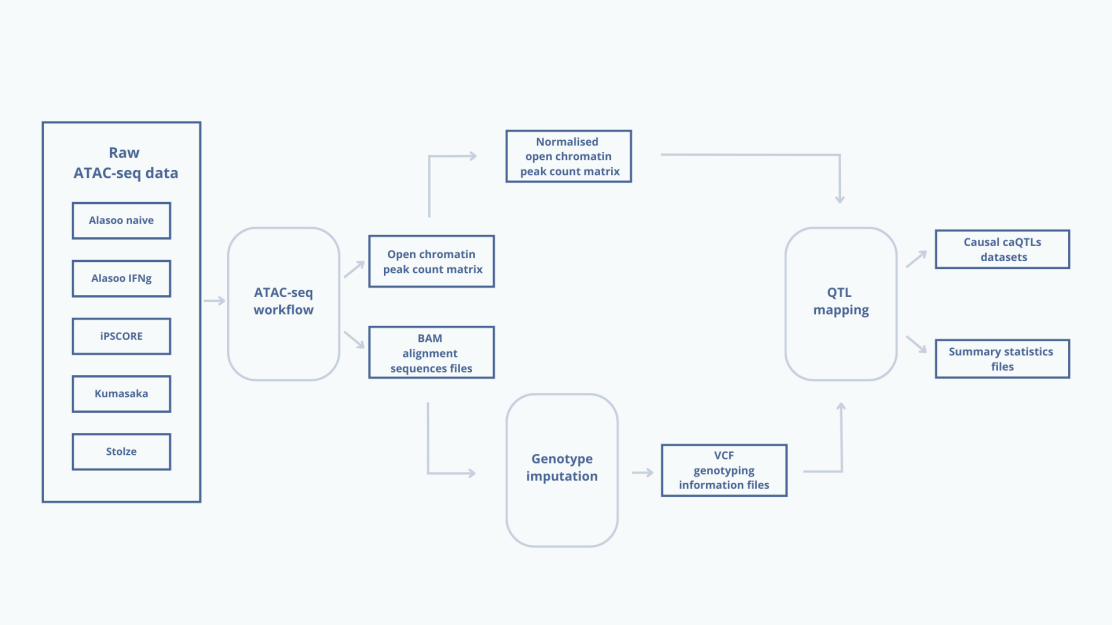


Figure 7. The overview of the steps from raw data to fine-mapping results. At first raw ATAC-seq data is the input for the Nextflow ATAC-seq workflow. This outputs two relevant files: BAM alignment sequences and open chromatin peak count matrix files. BAM files are input for the genotype imputation workflow which gives out genotyping information as a VCF file. Finally, VCF file and normalised count open chromatin peak count matrix are necessary input for QTL mapping step. As a result of the last workflow, summary statistics files and caQTL datasets are produced.

The QTL mapping process gives out summary statistics files but also credible sets of the SuSiE fine-mapping. One of the files is put together by merging the credible sets file with summary statistics. This file was used for the following thesis results

analysis. The overview of the sequential steps of the workflows between raw input data and fine-mapping results is in Figure 7. Most relevant information is the molecular trait, the SNP, unique SNP id (rsid), the relevant chromosome, the position of the SNP, reference and alternate allele, pip-value, p-value and z-score. The alternate allele is also the effect allele, meaning it is responsible for the association with the trait. The z-score shows the effect size and the direction of the association between the genetic variant and traits, which in this case are the chromatin peaks. Z-score near zero means no or very small effect. If the z-score is positive, each additional alternate allele increases the measure of the trait and if it is negative, it acts vice versa.

3.6 Enformer comparison

To validate how well the Enformer predicts, the results of QTL mapping were compared against the Enformer variant effect predictions. Those predictions were already provided by the authors of the Enformer and were publicly available on the internet[10]. In addition to the prediction, they had information about the SNP, position of it, the chromosome, reference and alternate allele. The prediction they provided was the variant effect score and the metric for it was called SNP activity difference score (SAD)[25]. SAD score around 0 shows little or no effect on the association. Positive and negative values work the same as they do for z-scores.

As the Enformer predictions were based on an older GRCh37 reference genome compared to the GRCh38 used in this work, then the positions of the variants from fine-mapping results had to be aligned with the Enformer variants' positions. This was done with an R package called MungeSumStats[21]. After aligning the positions, the variants from QTL mapping were mapped with the Enformer variants by linking them with the position, the chromosome and the unique id of the variants. Besides that, there were cases where the reference and alternate allele were switched and therefore also the variant effects had opposite values. Because of this, z-scores were multiplied with -1 to get the correct values.

As Enformer had predicted effect scores for only common variants and the reference panel used in this work's genotype imputation workflow had more variants than the reference panel used in Enformer's paper, then it decreased the size of the variants that could be compared in the end. Variants located on the sex chromosomes X and Y were left out from the comparison as there were no Enformer predictions for those.

Out of the 5313 human tracks the Enformer predicted, 684 were chromatin accessibility-related (DNase/ATAC) [25]. To show the performance of Enformer and have a valid comparison, relevant cell types were picked from the 684 tracks for each of the datasets

[10]https://console.cloud.google.com/storage/browser/dm-enformer/variant-scores/1000-genomes/enformer;tab=objects?prefix=forceOnObjectsSortingFiltering=false

[10]https://github.com/deepmind/deepmind-research/tree/master/enformer

[10]https://www.bioconductor.org/packages/release/bioc/html/MungeSumstats.html

compared with Enformer. Although, as Avsec et al. [25] mentioned themselves, then in several cases there was not a DNase sample that would perfectly match with the measured sample cell type. Therefore, the most closely matched sample had to be chosen.

3.7 Technology

To run all the workflows in this thesis, the high-performance computing (HPC) center[11] of the University of Tartu was used. All of the main workflows were run with Nextflow[12]. Nextflow is a workflow framework that simplifies putting together many tasks and in addition, allows for parallelisation of those tasks. The amount of time each of the workflows took varied from 3 hours to more than 24 hours depending on the workflow and the amount of data processed. Scripts used for the workflows were all Linux scripts, but scripts for preparing the inputs for the workflows, results analysis and visualisation, were written in two programming languages, Python and R. Relevant scripts associated with this thesis can be seen in the GitHub repository *kristiinakuningas/thesis*.

[12] https://hpc.ut.ee/
[12] https://nf-co.re/

4 Results

In the results section, we firstly give an overview of the QTL mapping results. In addition, we analyse the difference between the fine-mapping causal genetic variants' measured effect and Enformer's predicted effect. We show the results of the analysis and discuss about what could be the possible reasons for the results to be as they are.

4.1 QTL mapping

Statistically significant associations could be filtered out with the low false discovery rate (FDR 10%) filter. Despite the fact that four datasets had all initially around 440-450K unique genetic variants (see Table 2), then there is a clear difference between the number of statistically significant unique genetic variants and peaks remaining after applying the FDR filter. Kumasaka's dataset statistically significant associations made up around 5% of the initial associations, although for other datasets it was under 1%. This could be explained because of the bigger sample size compared to others.

Table 2. The overview of data before and after fine-mapping

	Dataset **Cell type** **Donors**	**Alasoo_naive** **macrophage_naive** **41**	**Alasoo_IFNg** **macrophage_IFNg** **42**	**Stolze** **EC** **44**	**iPSCORE** **iPSC** **77**	**Kumasaka** **LCL** **100**
Before SuSiE fine-mapping	Open chromatin peaks	569658	569658	231687	548504	513973
	Open chromatin peaks (FDR <0.1) / %	3196 / 0.6%	3118 / 0.6%	291 / 0.1%	2622 / 0.5%	20655 / 4%
	Genetic variants	439509	438183	207436	442639	438136
	Genetic variants (FDR <0.1) / %	3091 / 0.7%	3004 / 0.7%	288 / 0.1%	2554 / 0.6%	19134 / 4.4%
After SuSiE fine-mapping	Open chromatin peaks	909	850	45	1363	12614
	Open chromatin peaks (PIP >0.8) / %	77 / 8.5%	87 / 10%	3 / 7%	168 / 12%	1782 / 14%
	Genetic variants	32368	32060	1060	41366	357278
	Genetic variants (PIP >0.8) / %	78 / 0.2%	90 / 0.3%	3 / 0.3%	161 / 0.4%	1663 / 0.5%

However, if we compare the largest dataset (Kumasaka) with second-largest dataset (iPSCORE), then from the associations' distributions in Figure 8.A and Figure 8.B, it is obvious that the distributions are different between two biggest datasets and that is also the reason why there were more significant associations for Kumasaka. When Kumasaka's dataset has obviously more near-zero associations then for iPSCORE it is vice versa and the distribution is rather left-skewed indicating more non-significant associations. It is obvious from the Figure 8.C that Stolze's associations distribution is also left-skewed similarly to iPSCORE. All of the other have right-skewed distributions.

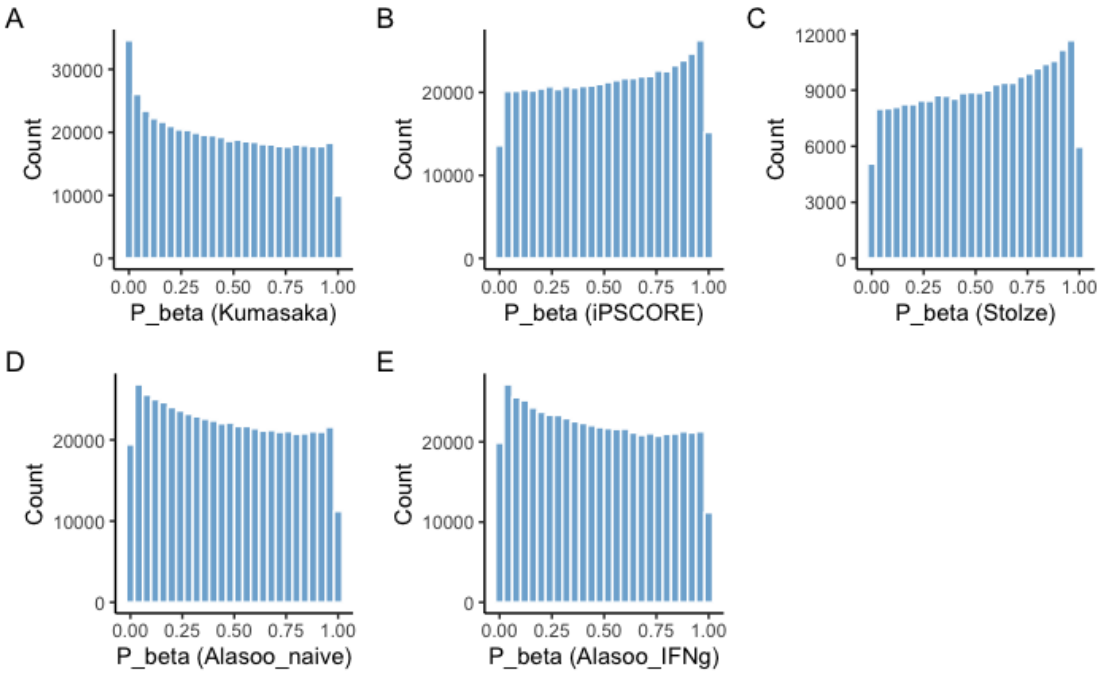


Figure 8. The distributions of statistically significant associations based on estimated empirical p-values based on beta distribution.

After fine-mapping, to be considered a potentially causative genetic variant for the trait in this thesis, a PIP threshold of 0.8 was used. PIP shows the strength of the evidence that this genetic variant is causal variant [15]. Therefore all the associations with PIP value over this threshold were considered as causal variants in this work and all below it were left out. As a result of this, one of the datasets (Stolze) was left with only 3 genetic variants. The biggest amount of potential causal variants was from Kumasaka's LCL data, where 1663 unique genetic variants remained. Despite the iPSCORE dataset having 77 samples, which is almost twice as much as Alasoo's both datasets, the amount of potential causal variants that remained was similar to that found in the Alasoo's datasets. The underlying reasons for this remain unknown for now, however, it could possibly be explained with the dataset's lower quality. The Figure 8.B distribution can also explain it a bit as it seems that there exist less significant associations for this dataset despite the bigger sample size.

It is evident based on Kumasaka's and iPSCORE's example that the more samples, the better the fine-mapping results and the more causal variants are determined. Another supporting fact for this statement is that iPSCORE had two times more causal variants than Alasoo's datasets had (see Table 2), although Alasoo's datasets had more statistically significant (FDR 10%) unique genetic variants. This also indicates that bigger sample sizes gives better fine-mapping results. In addition, it is probable that the quality of

the input ATAC-seq data or the differences in cell-type specific functions might also contribute to the explanations for the variability in the results.

As a result of fine-mapping, 5 different datasets of potentially causal genetic variants associated with chromatin accessibility were produced in this work. Different cell types covered were macrophage_naive, macrophage_IFNg, EC, LCL and iPSC. The number of potential causative variants varied from 3 to 1663 across different datasets. These datasets will give insight into how alterations in the genome in different cell types can influence chromatin accessibility that could in turn affect transcription factor binding and therefore gene expression in a way that can lead to a disease.

4.2 Comparison with Enformer

Datasets created by association mapping were then considered as validation datasets to evaluate Enformer's accuracy on chromatin accessibility predictions. However, not all of them could be used for it.

As could be seen in Table 2, the Stolze dataset had only 3 genetic variants that had a pip-value over 0.8. This could be because of the smaller size of the dataset, but also possibly because of the lower quality of the data, which both influence the results of fine-mapping [15,19]. As there were so few potentially causal variants after the fine-mapping, then this dataset was eliminated from the Enformer comparison.

Despite Alasoo's both datasets (naive and IFNg) having a considerable amount of potentially causative genetic variants after fine-mapping (PIP > 0.8), they were still left out of the Enformer comparison due to the fact that the number of variants decreased by aligning them with Enformer predictions common variants. Eventually, there were too few variants for a meaningful comparison. There did not exist a good track related to macrophages in Enformer's predictions either.

Table 3. The overview of data used in Enformer comparison analysis

Dataset, cell type Enformer, track	iPSCORE (iPSC) Enformer (iPSC)	iPSCORE (iPSC) Enformer (LCL)	Kumasaka (LCL) Enformer (LCL)	Kumasaka (LCL) Enformer (LCL)	Kumasaka (LCL) Enformer (iPSC)
Genetic variants	140	140	1430	1430	1430
Genetic variants (low SADs filtered out) / %	51 / 36%	11 / 8%	228 / 16%	178 / 12%	107 / 7%
Opposite values / %	37 / 26%	73 / 52%	429 / 30%	436 / 30%	588 / 41%
Opposite values (low SADs filtered out) / %	5 / 10%	-	8 / 4%	2 / 1%	-
Spearman correlation	0.47	-0.06	0.5	0.49	0.27
Spearman correlation (low SADs filtered out)	0.83	-	0.76	0.76	-

Therefore, Enformer predictions were compared only with Kumasaka and iPSCORE datasets, which both contained a considerable amount of potentially causal variants, especially Kumasaka (see Table 3). It is worth mentioning that they also had approximately

double as many samples as both Alasoo's and Stolze's datasets (77 and 100 donors against 41, 42 and 44 donors accordingly). This indicates that more samples produce better fine-mapping results and sample sizes under 70-80 are not suitable for this kind of QTL mapping.

To evaluate how Enformer performs, the z-score (measured) and SAD score (predicted) were compared. These scores both show genetic variant effect scores and therefore they could be comparable. However, the scales are different and that can also be seen in the graphs. To show and analyse how well the Enformer does, scatter plots of measured and predicted effects were created. In addition, Spearman correlation metric was used to show the correlation between the effects. Spearman correlation metric helps to measure the strength of the relationship between variables without needing the scales of the variables to be the same [59].

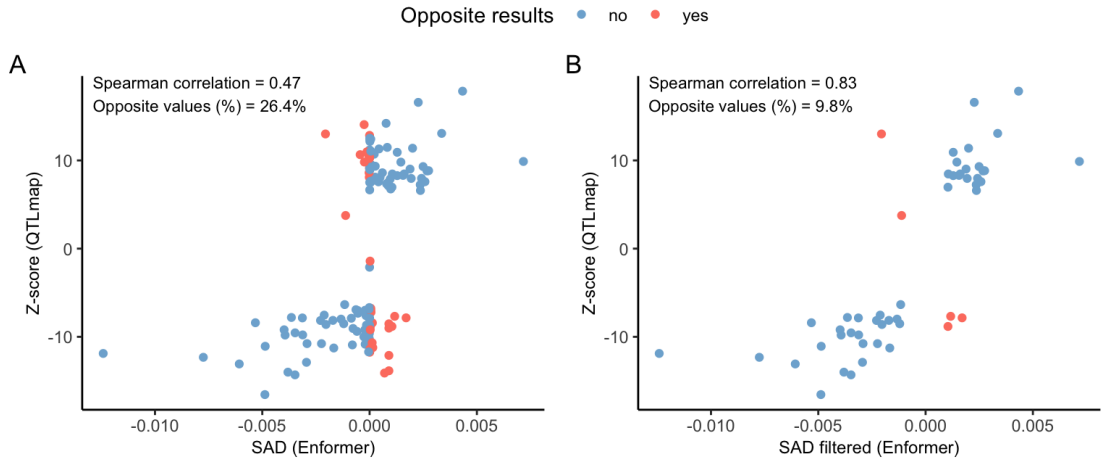
While looking at the results, it was clear that there existed very low SAD values that could possibly lead to two explanations. First scenario is that the variant from fine-mapping was causal for impacting chromatin accessibility, but Enformer predicted the variant to not have an effect and therefore Enformer did not predict correctly and thereby produced a false negative prediction. Another scenario is that the genetic variant measured as causal from fine-mapping was prioritised incorrectly and is not actually causal. Therefore, Enformer predicts correctly, but fine-mapping provides false-positive results. To analyse the results more thoroughly and to reach more conclusions, we additionally provided a comparison with values where low SAD scores were filtered out. SAD scores with absolute value less than 0.001 were considered as low SAD scores in this work.

To compare the Enformer results with the iPSCORE dataset, we chose one relevant prediction track from Enformer. It was described as "GM23338 male adult (53 years)", which represents iPSCs derived from fibroblasts and was therefore eligible for comparison with measured causal variants effects. Term "measured" means in this work the results derived from the fine-mapping analysis. We compared 140 genetic variants from iPSCORE fine-mapping results aligned with Enformer common variants. For the Enformer to predict well we would expect that the measured z-scores and predicted SAD scores would have the effect in the same direction (positive or negative) and have similar effect sizes as well.

In Figure 11, the left scatter plot (Figure 9.A) represents the measured effect and the Enformer's prediction on the effect of variants in iPSC. The right (Figure 9.B) represents the same plot as on the left but with the low SAD scores filtered out. The blue dots represent the genetic variants which measured and predicted effects in the same direction. In other words, the dots are blue if both Enformer prediction of the variant effects and measured effect were positive or both negative. A positive value of the effect score means an increase in chromatin accessibility and negative a decrease with respect to the alternative allele. Red dots, on the other hand, represent opposite directions of measured

and predicted genetic variant effects. For the Enformer to perform well, it is expected that there are no red dots or the number of those is as minimal as possible.

Based on Figure 9 filtered results on the right, we can say that Enformer did pretty well. We can see clearly that there's a clear linear trend between measured and predicted effects - if the measured effect increased, the Enformer's predicted effect increased and vice versa. 10% of the filtered variants had opposite effects. The Spearman correlation was 0.83 for the filtered values indicating a rather high correlation between the results. On the right, we can see the unfiltered results where around 64% out of the total 140 variants had very low SAD scores. The Spearman correlation for all the data points was 0.47.

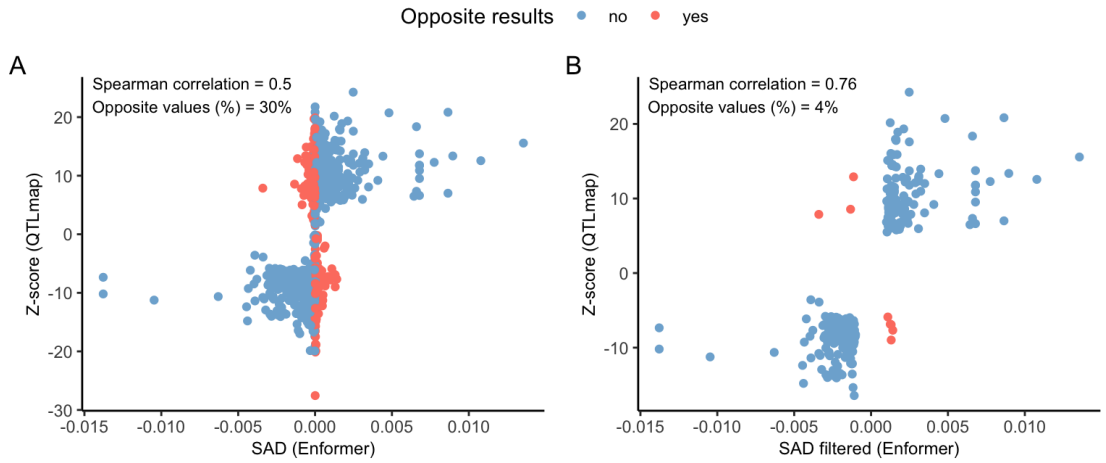


QTLMAP measured cell type = iPSC
ENFORMER prediction cell type = GM23338 male adult (53 years) (iPSC)

Figure 9. iPSCORE's causal genetic variants' measured effects on chromatin accessibility in comparison with Enformer's predicted effects in iPSC. Graph A represents the measured effect (z-score) on the y-axis and Enformer's predicted effect (SAD) on the x-axis. Graph B shows the same results but low SAD scores are filtered out. Blue dots on the graphs show genetic variants where the measured and predicted effects have the same direction. Red dots, on the other hand, represent the case where measured and predicted effect have opposite values.

For Kumasaka's dataset, 1430 variants were left for comparison after aligning the variants with Enformer's variants. Tracks that we analysed were GM12892 and GM19239, which both represent LCLs which is also the cell type of Kumasaka's dataset. Firstly, in Figure 10.B, we can again see how with filtered results there is a clear linear trend. The Spearman correlation of 0.76 indicates also a rather strong correlation. The portion of opposite results from filtered results was only 4%. Based on these filtered results, where only genetic variants with bigger effects are presented, we could say that Enformer

performs good. However, if we look at Figure 10.A, 84% of the 1430 unfiltered variants have very low predicted effect scores. The Spearman correlation for unfiltered results was 0.5. As discussed earlier, possible reasons for that many low SAD scores this could be that the Enformer produced false-negative predictions or Enformer predicted correctly but the genetic variant from fine-mapping results was not actually causal.



QTLMAP measured cell type = LCL
ENFORMER prediction cell type = GM12892 (LCL)

Figure 10. Kumasaka's causal genetic variants' measured effects on chromatin accessibility in comparison with Enformer's predicted effects in LCL. Graph A represents the measured effect (z-score) on the y-axis and Enformer's predicted effect (SAD) on the x-axis. Graph B shows the same results but low SAD scores are filtered out. Blue dots on the graphs show genetic variants where the measured and predicted effects have the same direction. Red dots, on the other hand, represent the case where measured and predicted effect have opposite values.

Another point of view on Enformer's performance based on Kumasaka's data was using the second track, GM19239. Here the filtered graph shows (see Figure 11.B) similarly to previous graphs a clear trend indicating it can predict several variants' effects well. In addition to the similar effect sizes, it also predicts the same direction. Spearman correlation on filtered results was 0.76 again. It has an even lower amount of opposite predictions (1% out of 178) than previous Kumasaka's example (4% out of 228). The portion of low SAD scores (see Figure 11.A), however, is even higher (88%). Spearman correlation on unfiltered results is 0.49.

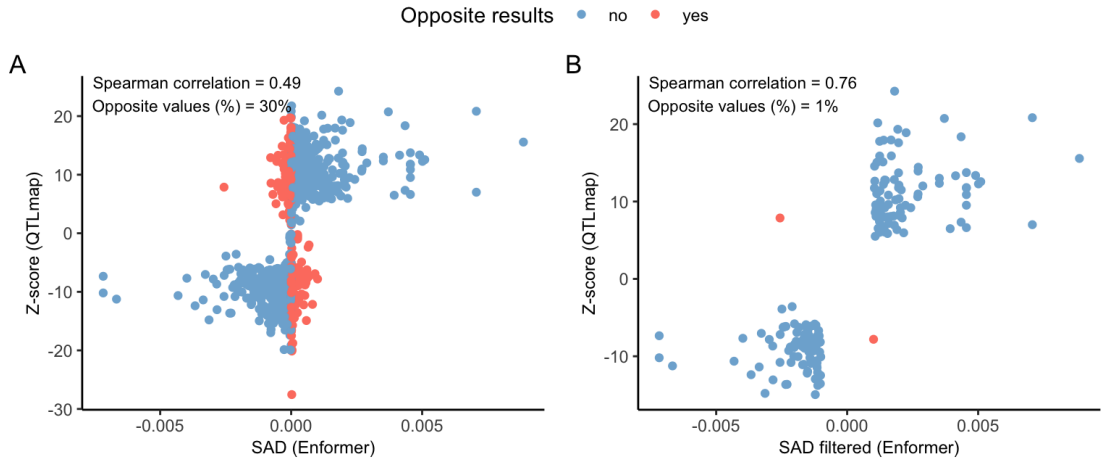
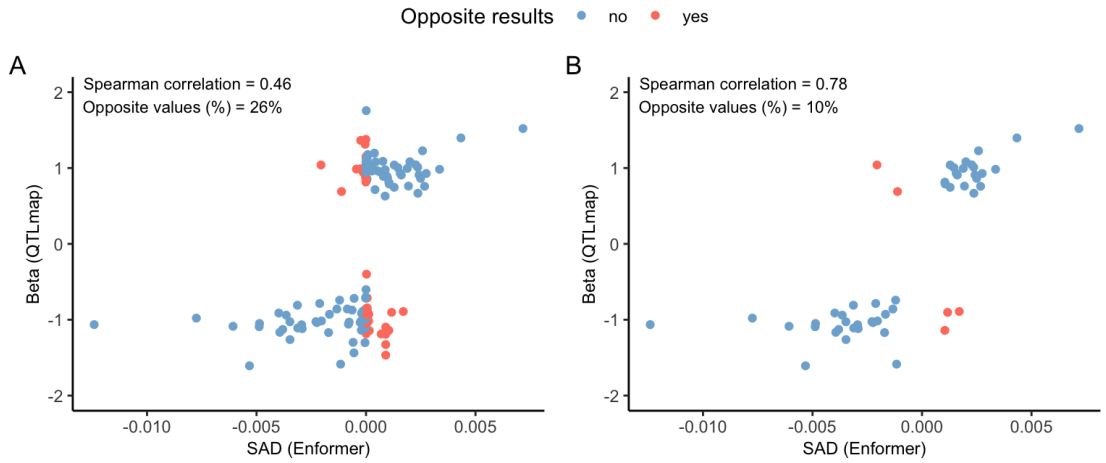


Figure 11. Kumasaka's causal genetic variants' measured effects on chromatin accessibility in comparison with Enformer's predicted effects in LCL. Graph A represents the measured effect (z-score) on the y-axis and Enformer's predicted effect (SAD) on the x-axis. Graph B shows the same results but low SAD scores are filtered out. Blue dots on the graphs show genetic variants where the measured and predicted effects have the same direction. Red dots, on the other hand, represent the case where measured and predicted effect have opposite values.

As we could see based on iPSCORE's and Kumasaka's examples, filtered results showed a rather good performance of Enformer. We could see that the Spearman correlation was a bit stronger for iPSCORE dataset (0.83 vs 0.76). One hypothesis to explain this is that from QTL mapping results we could determine that iPSCORE had less statistically significant genetic variants, meaning that the causal variants from fine-mapping could all be with higher effect on the trait and therefore it was easier for Enformer to predict. We can visualise the differences between the effect sizes between the datasets to test the hypothesis. For this we have to change the z-score to beta, because z-score could not be compared across different datasets as they have different scales then. With beta, which also shows the same effect as z-score, but is not standardised like z-score is, we could compare the two different datasets as they are on the same scale then. We looked for whether iPSCORE results had systematically higher beta values than Kumasaka's data or not. Higher values could support our hypothesis that it might have been easier for Enformer to predict.

If comparing Figure 12 and Figure 13 we cannot detect any signs that iPSCORE has higher beta values than Kumasaka. Therefore this would not explain our assumptions. However, one thing that can be noticed is that Kumasaka seems to have more near-

zero beta values if to compare Figure 12.A with Figure 13.A, but we cannot make any conclusions based on that. We also have to keep in mind that if data is more noisy, then beta values could decrease and therefore if these two datasets are of very different quality, then this comparison is not really valid. Therefore, this analysis step did not really give us any explanatory insight.



QTLMAP measured cell type = iPSC
ENFORMER prediction cell type = GM23338 male adult (53 years) (iPSC)

Figure 12. iPSCORE's causal genetic variants' measured effects on chromatin accessibility in comparison with Enformer's predicted effects in iPSC. Graph A represents the measured effect (beta-score) on the y-axis and Enformer's predicted effect (SAD) on the x-axis. Graph B shows the same results but low SAD scores are filtered out.

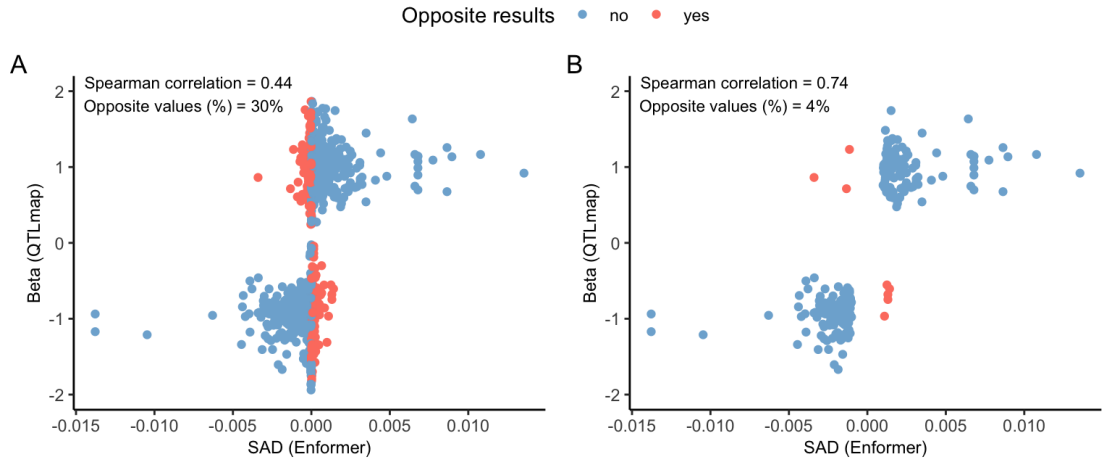
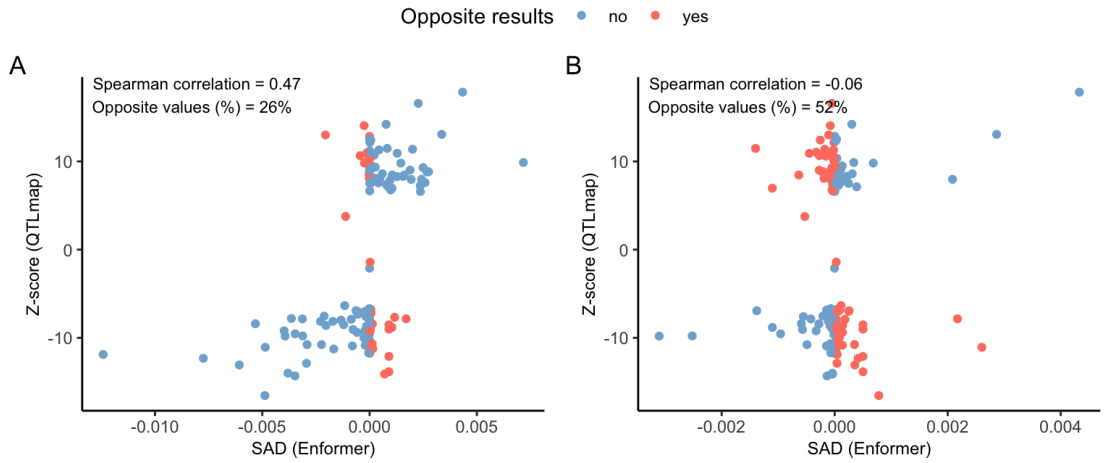


Figure 13. Kumasaka's causal genetic variants' measured effects on chromatin accessibility in comparison with Enformer's predicted effects in LCL. Graph A represents the measured effect (beta-score) on the y-axis and Enformer's predicted effect (SAD) on the x-axis. Graph B shows the same results but low SAD scores are filtered out.

Previously, we evaluated Enformer's predictions in the same cell types that the measured variants had. For example, measured genetic effect in LCL was compared with Enformer's prediction on a track directly related to LCL. In this case we expected the model to have linear same-directional results. Another possibility to validate Enformer's predictions is to analyse how Enformer performs if the cell types of measured and predicted variants are dissimilar, unlike previous examples. As chromatin accessibility is cell-specific and genetic variants can affect chromatin accessibility in one cell type but not in the other [31], then the variants' effects are expected to be different for different cells. If to compare measured effects from one cell type with Enformer's predictions from a different cell type, then we would expect there to be clearly more opposite effects on the variants. As LCLs and iPSCs are different cell types, then they were used against each other to see how would Enformer perform in this situation. In more detail, if Enformer predicted LCL variants' effects, then it was compared against measured iPSC cells variants' effects and vice versa.

Figure 14 represents the results of this experiment for measured iPSC causal variants and model predictions on the LCL track. Based on the visualisation it is obvious that the results do not match that well and there are more opposite values (red spots) as well. In addition, a clear trend has also disappeared. If we compare the Spearman correlations of the similar cell type comparison (Figure 14.A) vs the different cell type comparison (Figure 14.B), then they are 0.47 and -0.06 accordingly. This indicates that there is almost

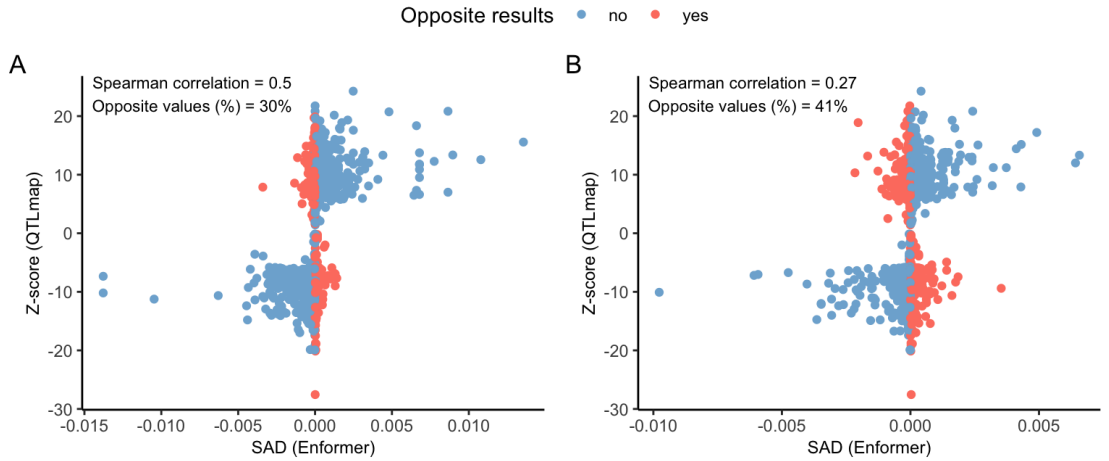
no correlation between the data points when compared cell types are not the same. This is characteristic of a well-performing model. In addition, we can see that the percentage of low SAD scores increased from 64% to 92%. This could also show that Enformer did not see any effect on these variants in that different cell type and therefore predicted correctly.



QTLMAP measured cell type = iPSC
ENFORMER prediction cell type = GM23338 male adult (53 years) (iPSC), GM12892 (LCL) (right)

Figure 14. iPSCORE's causal genetic variants' measured effects on chromatin accessibility in comparison with Enformer's predicted effects in the same cell type iPSC (A) and in a different cell type LCL (B). Graph A represents the measured effect (Z-score) on the y-axis and Enformer's predicted effect (SAD) on the x-axis in the iPSC cell type. Graph B shows the same as graph A but in a different cell type LCL.

Looking at the same analysis but on Kumasaka's data where the measured cell type is LCL and the predicted cell type is iPSC, we can see similar changes in the pattern as for the previous example (see Figure 14). A clear change in the trend can be detected in the visualisation of the results as it is not as obvious anymore in the right graph (see Figure 15.B). There are a lot more opposite values and it indicates the model is doing good. Spearman correlations drop from 0.49 (Figure 15.A) to 0.27 (Figure 15.B). In addition, the amount of low SAD values increase even more (84% vs 93%) with the dissimilar cell types.



QTLMAP measured cell type = LCL (left)
ENFORMER prediction cell type = GM12892 (LCL) (right), GM23338 male adult (53 years) (iPSC) (right)

Figure 15. Kumasaka's causal genetic variants' measured effects on chromatin accessibility in comparison with Enformer's predicted effects in the same cell type LCL (A) and in a different cell type iPSC (B). Graph A represents the measured effect (Z-score) on the y-axis and Enformer's predicted effect (SAD) on the x-axis in the LCL cell type. Graph B shows the same as graph A but in different cell type iPSC.

The total results in numbers from Enformer comparison analysis can be seen in Table 3. For those causal genetic variants, where Enformer predicts a larger effect on chromatin accessibility (low SAD scores filtered out), we can clearly see that the concordance between measured and predicted effects is pretty good. In addition, Spearman correlation for iPSCORE and Kumasaka in this case was 0.83 and 0.76 accordingly, which is quite a high result.

It also appears that Enformer predictions based on two cell types different from each other (iPSC vs LCL) lead to a smaller Spearman correlation value which could be expected because predictions of different cell types should be more different. For instance, for iPSCORE dataset, the Spearman correlation dropped from 0.47 to -0.06 if iPSC cell genetic variant effects were compared against LCL genetic variant predictions. In Kumasaka's case, it dropped from 0.49 to 0.27. In addition, the percentage of opposite values out of the total variants examined increased for both iPSCORE's and Kumasaka's data.

All in all, an initial overview of how Enformer predicts causal genetic variants' effect on chromatin accessibility has been shared. There are indications of Enformer predicting well, especially when it found a stronger effect (low SAD filtered out), as the data points of measured and predicted effects had linear trend between them, Spearman correlation values were high and there weren't that many opposite values either. In addition, based

on the experiments with different cell types (iPSC vs LCL; LCL vs iPSC), it was clear that results showed more opposite effects and lower Spearman correlation when cell types were different indicating Enformer's good performance.

However, for those genetic variants where Enformer predicts near-zero genetic variant effects, there could be two possible scenarios: Enformer gives false negative predictions or genetic variants from fine-mapping are not the correct ones. These are assumptions at the moment because based on the data we have here it is hard to distinguish what is the actual reason.

All in all, the first initial glimpse of how Enformer predicts chromatin accessibility has been shared. There are indications that Enformer performs quite well as the more confident predictions have a clear linear trend between the measured effects of causal variants. In addition, those variants have respectively high Spearman correlations of 0.76 and 0.83 for both LCLs and iPSCs. Furthermore, experiments with dissimilar cell types showed that the clear linear trend disappeared and it could be seen from especially iPSCORE dataset perspective. The correlation between the variants in those experiments also drops, which is expected. Although, it is clear that for a lot of genetic variants, Enformer predicts very low effect scores indicating that it is not confident in those predictions or it just predicts falsely.

Altogether, Enformer did pretty well especially when it predicted stronger effects. However, there were nuances related to low SAD scores that could not be explained at the moment with the data we have. Therefore, the performance of Enformer could be looked further into and tested on more cell types and more genetic variants from more data, but the first insight into how well Enformer predicts caQTL associations' effects has been provided.

5 Discussions

As a result of the SuSiE fine-mapping, 5 validation datasets containing information about potentially causal variants that could influence chromatin accessibility and therefore other regulatory functions were created. As chromatin accessibility is cell-type specific [31], then each of the datasets represents genetic variants and their associations with the phenotypic trait in each cell type separately. In addition to the information about the genetic variant id and the trait that it potentially affects, supplementary data about the variant and trait positions, relevant pip, p-values, effect sizes (z-score, beta) and more were provided in the dataset. Although ATAC-seq datasets used as an input for these datasets had been used earlier in studies [49-52], then they had not been used together in one work nor had they been processed with the same tools and workflows as used in this work.

QTL mapping and fine-mapping experiments showed that differences in sample sizes used in fine-mapping could lead to different quality results which has also been brought out in other studies [3,17,38]. It was clear that if the ATAC-seq dataset had under 50 samples, then it did not result in very many potential causal variants and the more samples were used, the more statistically significant associations were found and therefore also more causal variants in the end. The SNPs were considered causal in this thesis if they had a PIP value of over 0.8.

Although some studies such as Karollus et al. [33] and Sasse et al. [32] had already evaluated Enformer's performance on predicting the genetic variants' effects on phenotypic traits, then they did it from the gene expression perspective and on RNA-seq data. This thesis, on the other hand, provides insight into Enformer's execution on caQTLs based on validation datasets developed from ATAC-seq data QTL association mapping. This should provide an initial glimpse of how well Enformer performs and could be used to decide whether it can be possibly considered for future research.

5.1 Future enhancements

The selection of the ATAC-seq datasets for this work was based on the fact that there are not that many bigger ATAC-seq datasets available overall. If bigger ATAC-seq datasets were to become available, then they could definitely be applied to enhance the results of this work by allowing to detect more variant-trait associations across more cell types. Each additional unexplored dataset and cell type can lead to researchers one step closer to deciphering the *cis*-regulatory code.

What is more, datasets other than ATAC-seq could be used to perform more of these kind of association studies. For example ChIP-seq [61] data could be used to further detect genetic variants that could potentially link genetic variations in transcription factor binding sites to phenotypic traits or diseases, just like Baca et al. [39] did in their 2022 study where they found associations between cancer risk and TF binding loci.

Furthermore, RNA-seq data could be used from even more cell types to find more QTLs influencing gene expression.

From the perspective of Enformer's performance validation, additional bigger ATAC-seq datasets could lead to an even better understanding of how Enformer performs on predicting genetic variants associations with chromatin accessibility. In addition, the validation could also be done on ChIP-seq data while using Enformer's predictions on the ChIP tracks. Furthermore, some additional data to validate fine-mapping results could also provide us even better results from the validation.

6 Conclusion

In conclusion, genome-wide association studies have found several associations between genetic variants and diseases, although, in many cases, the underlying non-coding variants and their associations with regulatory functions that could affect the development of that certain disease remain unknown. Altogether, mapping genotypes to phenotypes is of high interest in genetics studies. Quantitative trait loci association mapping with fine-mapping has been a successful method to detect potentially causal genetic variants affecting chromatin accessibility. Deep learning models have also been a promising addition to the research field to better explain those complex aspects of non-coding genetic variants.

In this thesis, datasets of genetic variants that potentially have causal effects on chromatin accessibility in different cell contexts were created. To conduct those, data from 304 donors across 5 different cell types were used. The cells used were naive macrophages (macrophage_naive), macrophages stimulated with cytokine interferon-γ stimuli (macrophage_IFNg), endothelial cells (EC), lymphoblastoid cell lines (LCL) and induced pluripotent stem cells (iPSC). CaQTL datasets of sizes differing from 3 to 1663 were created. In addition to providing the ids of the genetic variants and the open chromatin regions, additional metadata such as the positions of relevant variants and traits, fine-mapping PIP values, p-values and effect score were provided.

In addition, the generated datasets were used as validation data for evaluating a state-of-the-art deep learning model Enformer's predictions on associations' effects between common genetic variants and chromatin accessibility. The results showed some clear positive patterns between experimentally measured and predicted effects indicating that Enformer could predict somewhat well, especially on the variants that the Enformer predicted confidently. In addition, when comparing the model's predictions on one cell type that is different from the measured cell type, then it did show less correlation between the predicted and measured effects, which is also expected for a well-performing model. However, it still provided a lot of near-zero values indicating there were variants where the model could not predict the effects well or at all.

To conclude, this thesis provided a base knowledge of genetics to allow a reader not related to the field to comprehend the topic better. All the underlying theories and relevant terms used in the experiments were described and explained in this work. In addition to the background knowledge, readers were provided with insights into the main steps of this work: open chromatin data processing, alignment of the data to the reference genome, genotype imputation, QTL mapping including SuSiE fine-mapping and Enformer evaluation.

All in all, as a result of this thesis, beneficial sources for further research in the field have been provided. The Enformer performance evaluation on ATAC-seq data can give a first insight into how Enformer predicts overall on ATAC-seq data and this could be looked further into to possibly implement this in research more. What is more, the

created datasets can bring more perspective on whether and how chromatin accessibility has influenced the development of some diseases or how to develop new treatments for those. Although, further studies have to be conducted to make final conclusions on whether the potentially causal variant actually affected the trait or disease. All of the results from this thesis could be used as a piece of additional knowledge in this field.

References

[1] Heinaru A. Geneetika. Õpik kõrgkoolile. Tartu: Tartu Ülikooli Kirjastus. 2012. http://geneetika.ee/geneetika-opik/ (09.05.2023)
[2] Watson J.D., Crick F.H.. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* , 1953, No. 171, pp 737-738.
[3] Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals. Washington (DC): Genetic Alliance. 2009.
https://www.ncbi.nlm.nih.gov/books/NBK115568/. (09.05.2023)
[5] Clancy S., Brown W. Translation: DNA to mRNA to Protein. *Nature Education*, 2008, No 1, pp 101. https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/
[6] Luger K., Mäder A., Richmond R. et al. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 1997, No. 389, pp 251–260. https://doi.org/10.1038/38444
[7] Aguet F., Alasoo K., Li Y.I. et al. Molecular quantitative trait loci. *Nat Rev Methods Primers*, 2023, 3. https://doi.org/10.1038/s43586-022-00188-6
[8] Howie B.N., Donnelly P., Marchini J.. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, 2009, 5(6). https://doi.org/10.1371/journal.pgen.1000529.
[9] Uffelmann E., Huang Q.Q., Munung N.S. et al. Genome-wide association studies. *Nat Rev Methods Primers* , 2021, No. 1, pp 59. https://doi.org/10.1038/s43586-021-00056-9
[10] Matharu N., Ahituv N. Modulating gene regulation to treat genetic disorders. *Nat Rev Drug Discov*, 2020, No 19, pp 757–775. https://doi.org/10.1038/s41573-020-0083-7
[11] Wittkopp P., Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet*, 2012, No 13, pp 59–69. https://doi.org/10.1038/nrg3095
[12] Myles S., Peiffer J., Brown P.J. et al. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. *The Plant Cell*, 2009, Vol 21, No 8, pp 2194–2202. https://doi.org/10.1105/tpc.109.068437
[15] Wang G., Sarkar A., Carbonetto P., Stephens M.. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology,* 2020, Vol 82, No 5, pp 1273–1300. https://doi.org/10.1111/rssb.12388
[16] Schaid D.J., Chen W., Larson N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*, 2018 No 19, pp 491–504. https://doi.org/10.1038/s41576-018-0016-z
[20] Hutchinson A, Watson H, Wallace C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLOS Computational Biology*, 2020, Vol 16, No 4. https://doi.org/10.1371/journal.pcbi.1007829
[21] Murphy A.E., Schilder B.M., Skene, N.G. MungeSumstats: a Bioconductor package

for the standardization and quality control of many GWAS summary statistics. *Bioinformatics*, 2021, Vol 37, No 23, pp 4593–4596. https://doi.org/10.1093/bioinformatics/btab665
[22] Boer C.G.B., Taipale J. Hold out the genome: A roadmap to solving the cis-regulatory code. *bioRxiv*, 2023. https://doi.org/10.1101/2023.04.20.537701
[24] Kim S. Wysocka J. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell,*, 2023, Vol 83, No 3, pp 373-392.
https://doi.org/10.1016/j.molcel.2022.12.032.
[25] Avsec, Ž., Agarwal, V., Visentin, D. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* , 2021, Vol 18, pp 1196–1203. https://doi.org/10.1038/s41592-021-01252-x
[26] Shlyueva D., Stampfel G., Stark A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*, 2014, Vol 15, pp 272–286.
https://doi.org/10.1038/nrg3682 Enhancers
[27] Zhou J., Theesfeld C.L., Yao K. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*, 2018, Vol 50, pp 1171–1179. https://doi.org/10.1038/s41588-018-0160-6
[28] Kelley D.R. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology*, 2020, Vol 16, No 7.
[30] https://www.biorxiv.org/content/10.1101/737981v3.full cis-regulatory DL
[31] Liang, D., Elwell, A.L., Aygün, N. et al. Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation. *Nat Neurosci*, 2021, Vol 24, pp 941–953. https://doi.org/10.1038/s41593-021-00858-w
[32] Sasse A., Ng B., Spiro, A., et al. How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks? *bioRxiv*, 2023. https://doi.org/10.1101/2023.03.16.532969
[33] Karollus A., Mauermeier T., Gagneur J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol*, 2023, Vol 24, No 56. https://doi.org/10.1186/s13059-023-02899-9
[35] Whalen S., Pollard K.S. Reply to 'Inflated performance measures in enhancer–promoter interaction-prediction methods'. *Nat Genet*, 2019, Vol 51, pp 1198–1200.
https://doi.org/10.1038/s41588-019-0473-0
[36] de Koning A.P.J., Gu W., Castoe T.A., Batzer M.A., Pollock D.D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLOS Genetics, 2011, Vol 7, No 12. https://doi.org/10.1371/journal.pgen.1002384
[37] Wong K.H.Y., Ma W., Wei CY. et al. Towards a reference genome that captures global genetic diversity. *Nat Commun*, 2020, 11, pp 5482. https://doi.org/10.1038/s41467-020-19311-w
[38] Stegle O., Parts L., Durbin R., Winn J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLOS Computational Biology*,2010, Vol 6, No 5.

https://doi.org/10.1371/journal.pcbi.1000770 sample size
[39] Baca S.C., Singler C., Zacharia S. et al. Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. *Nat Genet*, 2022, Vol 54, pp 1364–1375. https://doi.org/10.1038/s41588-022-01168-y
[40] Wang Y., Gao H., Wang F. et al. Dynamic changes in chromatin accessibility are associated with the atherogenic transitioning of vascular smooth muscle cells, *Cardiovascular Research,*, 2022, Vol 118, No 13, pp 2792–2804. https://doi.org/10.1093/cvr/cvab347
[41] Buenrostro, J., Giresi, P., Zaba, L. et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218 (2013). https://doi.org/10.1038/nmeth.2688
[42] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3627383/ DNASE
[43] MNASE https://pubmed.ncbi.nlm.nih.gov/17392789/
[44]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1891346/ FAIRE-seq
[46] Tsompana M., Buck M.J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin*, 2014, Vol 7, No 33.
https://doi.org/10.1186/1756-8935-7-33
[47] Yan, F., Powell, D.R., Curtis, D.J. et al. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol 21, 22 (2020). https://doi.org/10.1186/s13059-020-1929-3
[48] Elhelu M.A.. The role of macrophages in immunology. *J Natl Med Assoc*, 1983, Vol. 75, No 3, pp 314-7. PMID: 6343621; PMCID: PMC2561478.
https://pubmed.ncbi.nlm.nih.gov/6343621/ (09.05.2023)
[49] Alasoo, K., Rodrigues, J., Mukhopadhyay, S. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat Genet 50, 424–431 (2018). https://doi.org/10.1038/s41588-018-0046-7
[50] Stolze L.K., Conklin A.C. et al. Systems Genetics in Human Endothelial Cells Identifies Non-coding Variants Modifying Enhancers, Expression, and Complex Disease Traits. The American Journal of Human Genetics, 2020, Vol 106, No 6, pp 748-763. https://doi.org/10.1016/j.ajhg.2020.04.008.
[51] Kumasaka, N., Knights, A.J. Gaffney, D.J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nat Genet 51, 128–137 (2019). https://doi.org/10.1038/s41588-018-0278-6
[52] Omi, N., Tokuda, Y., Ikeda, Y. et al. Efficient and reliable establishment of lymphoblastoid cell lines by Epstein-Barr virus transformation from a limited amount of peripheral blood. Sci Rep 7, 43833 (2017). https://doi.org/10.1038/srep43833
[53] Greenwald, W.W., Li, H., Benaglio, P. et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. Nat Commun 10, 1054 (2019). https://doi.org/10.1038/s41467-019-08940-5
[54] Cock P.J.A, Fields C.J., Goto N. et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*,

2010, Vol 38, No 6, pp 1767–1771, https://doi.org/10.1093/nar/gkp1137
[55] Patel H., Espinosa-Carrasco J., Ewels P. et al. nf-core/atacseq: nf-core/atacseq v2.0 - Iron Iguana (2.0). Zenodo, 2022. https://doi.org/10.5281/zenodo.7384115
[57] Kolberg P. Ekspressiooni kvantitatiivsete tunnuste loouste analüüs üksikraku RNA sekveneerimisandmetes. 2023.
[58] Beasley, T.M., Erickson, S. Allison, D.B. Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited?. Behav Genet 39, 580–595 (2009). https://doi.org/10.1007/s10519-009-9281-0
[59] Hauke J., Kossowski T.. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. Quaestiones Geographicae. 2011;30(2): 87-93. https://doi.org/10.2478/v10117-011-0021-1
[61] Park, P. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 2009 10, pp 669–680. https://doi.org/10.1038/nrg2641
[63] Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J. et al. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet* , 2021, Vol 53, 120–126. https://doi.org/10.1038/s41588-020-00756-0
[64] Wang D., Wu X., Jiang G. et al. Systematic analysis of the effects of genetic variants on chromatin accessibility to decipher functional variants in non-coding regions. *Frontiers in Oncology*, 2022, Vol 12. https://doi.org/10.3389/fonc.2022.1035855
[65] Marchini, J., Howie, B. Genotype imputation for genome-wide association studies. Nat Rev Genet 11, 499–511 (2010). https://doi.org/10.1038/nrg2796
[66] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945 (2004). https://doi.org/10.1038/nature03001
[67] Zhen, Y., Andolfatto, P. (2012). Methods to Detect Selection on Noncoding DNA. In: Anisimova, M. (eds) Evolutionary Genomics. Methods in Molecular Biology, vol 856. Humana Press. https://doi.org/10.1007/978-1-61779-585-5_6
[68] Palazzo A.F., Gregory T.R. The Case for Junk DNA. *PLOS Genetics*, 2014, Vol 10, No 5. https://doi.org/10.1371/journal.pgen.1004351
[69] Simna S.P., Han Zongchao. Prospects of Non-Coding Elements in Genomic DNA Based Gene Therapy. *Current Gene Therapy*, 2022, Vol 22(2). https://dx.doi.org/10.2174/1566523221666210419090357
[70] Rye C., Wise R., Jurukovski V. et al. 2016. Biology. Texas: OpenStax. https://openstax.org/books/biology/pages/1-introduction (09.05.2023)
[71] Alberts B., Johnson A., Lewis J., et al. Molecular Biology of the Cell. 4th edition. Chromosomal DNA and Its Packaging in the Chromatin Fiber. New York: Garland Science. 2002. https://www.ncbi.nlm.nih.gov/books/NBK26834/ (09.05.2023)
[72] Simpson B., Tupper C., Al Aboud N.M.. Genetics, DNA Packaging. 2022. Treasure Island (FL): StatPearls Publishing. https://pubmed.ncbi.nlm.nih.gov/30480946/
[73] Bhangale T.R, Rieder M.J., Livingston R.J. et al. Comprehensive identifica-

tion and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Human Molecular Genetics*, 2005, Vol 14, No 1, pp 59–69, https://doi.org/10.1093/hmg/ddi006
[74] Miles C., Wayne M. Quantitative trait locus (QTL) analysis. Nature Education, 2008, Vol 1, No 1, pp 208. https://www.nature.com/scitable/topicpage/quantitative-trait-locus-qtl-analysis-53904/ (09.05.2023)
[75] Schneider V.A., Graves-Lindsay T., Howe K. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*, 2017, Vol 27, pp 849-864. https://genome.cshlp.org/content/27/5/849
[76] Formenti G., Theissinger K., Fernandes C. et al. The era of reference genomes in conservation genomics. *Trends in Ecology Evolution*, 2022, Vol 37, No 3, pp 197-202. https://doi.org/10.1016/j.tree.2021.11.008
[78] Deng T., Zhang P., Garrick D. et al. Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. *Frontiers in Genetics*, 2022, Vol 12. https://doi.org/10.3389/fgene.2021.704118
[80] Eicher, J. D., Landowski, C., Stackhouse, B. et al. GRASP v2.0: An update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Research*, 2015, Vol 43, pp D799-D804. https://doi.org/10.1093/nar/gku1202
[81] Single Nucleotide Polymorphism (SNP). Encyclopedia of Public Health. Ed. by W. Kirc. 2008. Dordrecht: Springer, pp 1305.
https://doi.org/10.1007/978-1-4020-5614-7_3214.
[83] Ashcraft C.W. Genetic variability by design. *Journal of Creation*, 2004, Vol 18, No 2., pp 98-104. https://creation.com/images/pdfs/tj/j18_2/j18_2_98-104a.pdf (09.05.2023)
[84] Chial H. DNA sequencing technologies key to the Human Genome Project. *Nature Education,* 2008, Vol. 1, Issue 1, p. 219. https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/ (09.05.2023).
[85] Phillips T. Regulation of transcription and gene expression in eukaryotes. *Nature Education*, 2008, Vol 1(1), pp 199. https://www.nature.com/scitable/topicpage/regulation-of-transcription-and-gene-expression-in-1086/ (09.05.2023)

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Kristiina Kuningas**,

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,

   Estimating Concordance Between Measured and Predicted Genetic Variant Effects on Chromatin Accessibility,

   supervised by Kaur Alasoo.

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.

Kristiina Kuningas
09/05/2023