University of Tartu

Faculty of Science and Technology

Institute of Technology

Nihat Aliyev

Developing a computational workflow for eQTL analysis on the X chromosome

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor:

PhD Kaur Alasoo

Tartu 2021

Abstract

Developing a computational workflow for eQTL analysis on the X chromosome

Despite advances in sequencing technology and computational biology which led to identifying underlying causes for complex traits, utilization of X chromosome data lags behind the autosomes. This can be attributed to the inherent complexities of analyzing X chromosome data and extra data processing steps needed before the analysis. The aim of this thesis was to develop a computational workflow for the inclusion of X chromosome analysis and improve the shortcomings in order to supplement the existing eQTL analysis methods. We demonstrated that after adjustment of X chromosome dosage differences between females and males, existing workflows can be used to uncover potential causal variants for complex traits and diseases. Using RNA-seq data from human lymphoblastoma cell lines obtained from GEUVADIS project we performed statistical fine mapping and colocalization analysis with external databases. Results show significant associations of PLP2 gene with respiratory and cardiovascular functions.

CERCS:

B220 Genetics, Cytogenetics

P170 Computer science, numerical analysis, systems, control

Keywords:

GWAS, eQTL, X chromosome, Fine mapping, BCFtools

Kokkuvõte

X kromosoomil põhineva eQTL analüüsi arvutusliku töövoo väljatöötamine

Hoolimata edasiminekutest DNA sekveneerimise tehnoloogias ja arvutusbioloogias, mille abil on identifitseeritud kompleksete tunnuste põhjused, on X kromosoomi data analüüs autosoomidest maha jäänud. Selle põhjuseks võib tuua lisasammud X kromosoomi data töötlemisel enne analüüsimist ning X kromosoomi data analüüsimise keerukuse. Antud töö eesmärk oli täiustada olemasolevaid eQTL analüüsimeetodeid, arendades X kromosoomi kaasav arvutuslik töövoog ja parandada puudusi. Me näitasime, et peale X kromosoomi doosi erinevuse korrigeerimist meeste ja naiste vahel, on võimalik kasutada olemasolevaid meetodeid potentsiaalsete kompleksete tunnuste ja haiguste põhjuslike variantide tuvastamiseks. GEUDAVIS projektist saadud inimese lümfoblastoma rakuliini RNA-seq datat kastades sooritasime statistilise peen kaardistamise ning kolokalatsiooni analüüsi väliste databaaside abil. Tulemused näitavad statistiliselt olulisi seoseid PLP2 geeni ning respiratoorsete ja kardiovaskulatoorsete funktsioonidega.

CERCS:

B220 Geneetika, tsütogeneetika

P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Märksõnad:

GWAS, eQTL, X kromosoom, Peen kaardistamise, BCFtools

Contents

List of Figures

List of Tables

Abbreviations, definitions

eQTL - expression quantitative trait loci

GWAS - Genome-Wide Association Studies

OMIM - Online Mendelian Inheritance in Man

SNP - single nucleotide polymorphism

VCF - Variant Call Format

Introduction

Advances in molecular genetics led to the fast-paced identification of genes that are associated with human diseases. Traditionally research focused on single gene single disease model, where finding the gene responsible for abnormal phenotype was the goal. In fact, many well-studied diseases are transmitted in various monogenic (Mendelian) ways. Cystic fibrosis is caused by an autosomal recessive mutation in the CTFR gene, sickle-cell anemia is associated with an autosomal dominant mutation in the beta hemoglobin gene and an X-linked mutation in the gene responsible for the production of dystrophin leads to Duchenne syndrome. [1] Nevertheless, most traits are thought to be complex. Complex phenotypes arise from the accumulation of individually small effects of multiple genes and environmental factors. This is one of the main drivers of variation in population but also serves as a feedback mechanism to diminish adverse effects in the case of mutation in genes associated with the trait. [2]

The majority of the disorders arise when regulatory mechanisms can not compensate for the change (loss) of function caused by mutations. Most prevalent diseases such as diabetes, heart diseases, obesity, etc. are classified as complex diseases. [3] The very nature of complex traits makes it difficult to find associated regions in the genome that would explain the phenotype. The most resorted method for discovering these associations is Genome-Wide Association Studies (GWAS). In spite of generating a large amount of data, GWAS has its shortcomings which I will discuss further. Additionally, the existence of a certain variant in the genome does not always mean it will have any observable effect on the phenotype of interest. Expression of the variant or effect of the variant on the expression of another locus usually determines the outcome. Hence, eQTL analysis which focuses on the variance of gene expression is often performed after GWAS.

Despite the advances in computational biology, analyzing sex chromosomes remains challenging due to experimental constraints and the additional effort needed for data processing. This has resulted in the omittance of sex chromosomes from GWAS and post GWAS studies.

[4] It is crucial that we develop analysis methods that include sex chromosomes, especially the X chromosome that has been shown to carry loci associated with many known human diseases. This work aims to supplement existing eQTL analysis methods by developing a computational workflow that includes X chromosome analysis and improves on inherent shortcomings such as differences between males and females in the dataset.

The thesis is structured in the following way. Chapter 1 describes the literature on the topic and summarizes the problem in hand. Chapter 2 discusses the methods used in the study and brings out the results. The results are further discussed and summarized in Chapter 3.

1 Literature review

1.1 GWAS

First introduced in 2005, Genome-Wide Association Studies (GWAS) in which a collection of genetic variants are tested to link genotype to the observed phenotype, has revolutionized the study of complex disorders. [5] As the cost of genome sequencing came down exponentially, the number of GWAS published increased rapidly. [6] Benefits that came from it include the discovery of new drug targets, estimating disease susceptibility, and practical applications in personalized medicine fields such as adjusting the dosage of administered drugs based on the patient's genotype. [7] GWAS can also find associations with low frequency and rare variants. Predictably, the number of associations found significantly increases with the number of samples analyzed. This inevitably increases the cost associated with GWAS. As a result, most studies make use of genome-wide single nucleotide polymorphism (SNP) arrays and subsequent statistical imputation of unobserved genotypes using the reference panel. Moreover, GWAS only explains a fraction of the heritability of complex traits. [8] Most GWAS hits pinpoint to noncoding regions of the genome hence making it challenging to find the causal variant, hence the gene of interest. [9] Producing multiple association hits with a given trait further increases the challenge of identifying the causal variant. Considering all the limitations of GWAS focus has been shifting to post-GWAS research to further illuminate the genetic mechanism behind the complex diseases.

1.2 eQTL

With the increased availability of transcriptomic data which catalogs the mRNA levels in different cell and tissue types, eQTL analysis has been proposed to be the next step after GWAS. eQTL is defined as a locus that explains variance in expression levels of a gene. An SNP that

has been found to be associated with a trait in GWAS has a 3 times higher chance of being associated with gene expression, hence to be an eQTL [10]. Based on the enrichment of associated SNPs in different tissue and cell types eQTL analysis also allows identifying the causal cell type of the complex disease. Hu *et al.*. demonstrated that CD4+ T-cells are causal cell types for rheumatoid arthritis while B-cells are related to Lupus Erythematosus [11]. Fundamentally, eQTL analysis bridges the gap between the genetic variation and disease by profiling the intermediate phenotypes in the shape of SNP $\longrightarrow$ gene expression $\longrightarrow$ trait (Figure 1.1).

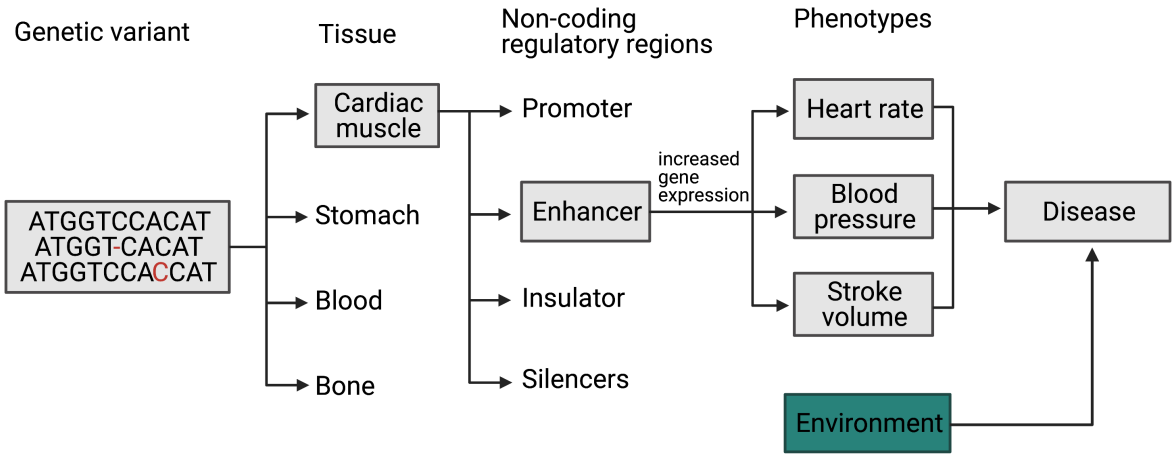

Figure 1.1: Flow from genetic variation to a trait. Variation in genetic code shows its effect on a trait through many intermediate phenotypes. This figure summarizes the road from Single Nucleotide Polymorphism to the hypothetical disease state.

eQTLs are classified as cis-eQTLs that act on local genes (+- 1MBP <) and trans-eQTLs that act on distant genes. Cis-regulatory variance derives from variation in regulatory regions where transcription factor binds while trans variation is generally the result of variation in transcription factor itself. Research suggests that most genes are regulated by cis-eQTLs and they tend to have larger effect sizes [12]. In the most simplistic way, eQTL analysis discovers the linear regression equation that relates the expression to the genotype. Using the equation of

$$Y_i = B_0 + B_1 X_i + \varepsilon_i \tag{1.1}$$

we can estimate a phenotype Y from the expression level X while accounting for basal expres-

sion of B_0, slope B_1 and error ε_i (Figure 1.2).

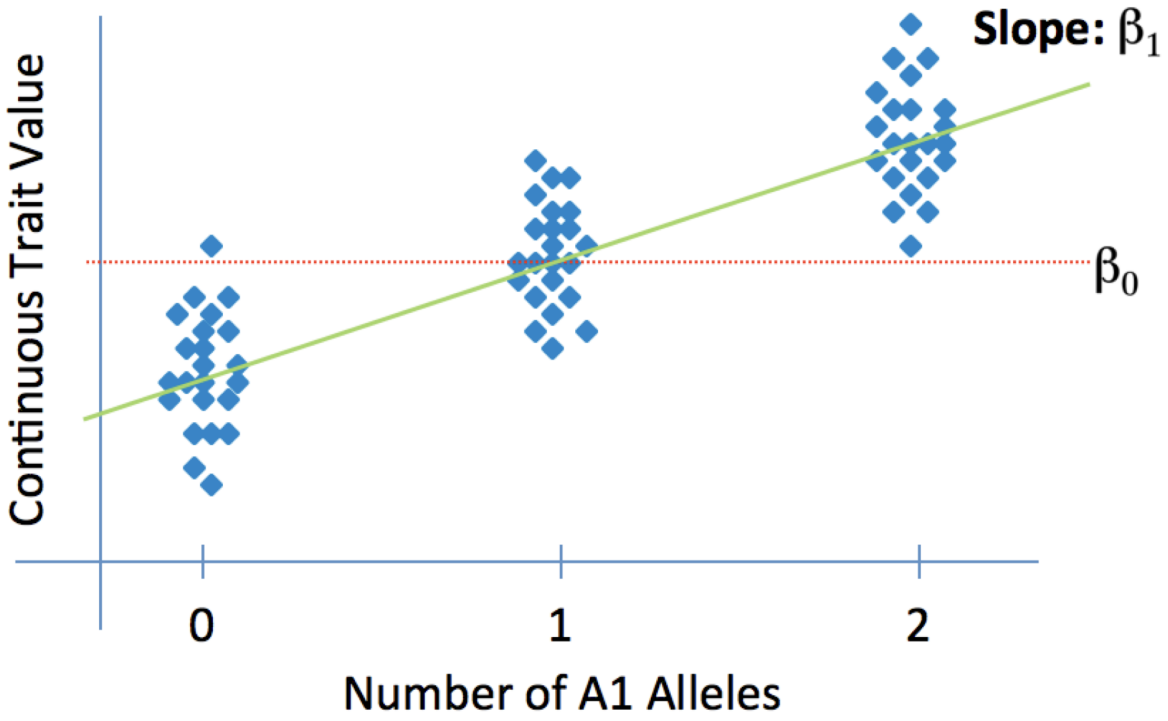

Figure 1.2: Linear regression graph of eQTL [13]. Trait value associated with variant increases proportionally to the number of alles of interest the individual possesses.

1.3 Genotype imputation

Although the cost of whole-genome sequencing has decreased many folds in recent years, it is yet not feasible to sequence thousands of people for analysis. Instead, researchers use SNP arrays with the size of 100000 to 1000000 variants. Considering the fact that more than 10 million estimated genetic variants exist, a typical study only covers a small fraction of the genome. This limited data is still helpful and can uncover many associations. A 2006 paper by Burdick *et al.* suggested a method for inferring the rest of the missing data computationally [14]. Now known as genotype imputation, this method makes use of a reference panel of haplotypes or genotypes and allows to evaluate associations of SNPs that are not directly genotyped. The main principle of imputations is that stretches of chromosomes are genetically linked and inherited together. This assumption holds for related and unrelated individuals with a significant difference being the much shorter shared haplotype stretch for unrelated individuals. [15] After the haplotype is identified by genotyped SNPs missing variants can be filled with different algorithmic approaches such as heuristic, expectation-maximization, or more complicated Markovian

coalescent models. Figure 1.3 depicts the simplified representation of genotype imputation.

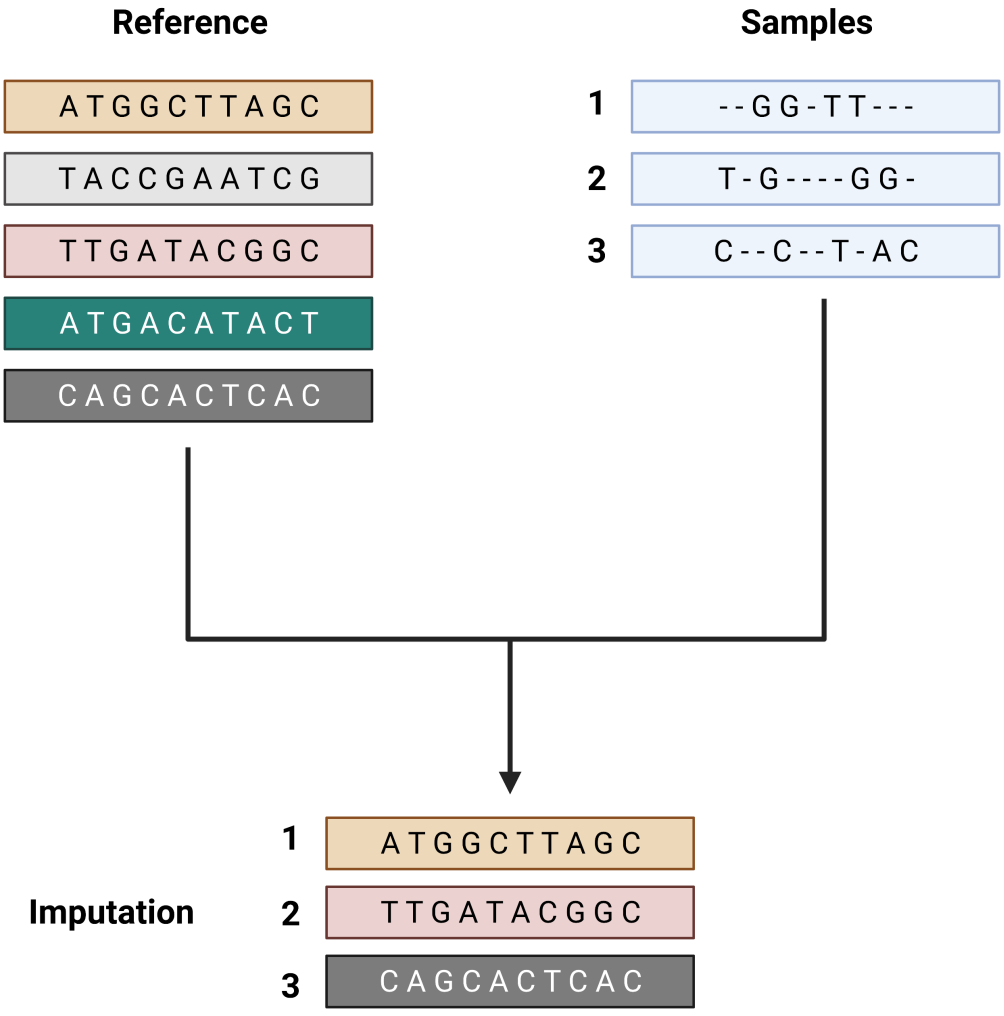

Figure 1.3: Genotype imputation. Graphical representation of a hypothetical imputation of haplotypes froma reference panel

There are two main classes of genotype imputation tools. The first class includes tools that use all observed variants to impute missing SNPs and are resource-intensive. IMPUTE [16], MACH [17], and fastPHASE/BIMBAM [18] are examples of this class. The second more computationally efficient software only uses markers near the imputed genotype to make a prediction. Well-known examples include PLINK [19], TUNA [20], and BEAGLE [21]. These tools provide remarkable accuracy of over 90 percent [15]. Whenever a variant can not be imputed with high confidence, most of the mentioned tools provide probability scores for the identity of the genotype. Obtained partial information can still be effectively used in association

analysis.

There are several use cases of genotype imputation. It increases the power of GWAS studies. On average 10% more peaks can be obtained on loci of interest after imputation. [22] It can also accelerate genetic fine-mapping research. After obtaining an association signal, genotype imputation allows us to zoom in and test for association in nearby SNPs. This procedure aids in identifying the potential causal variants. Furthermore, genotype imputation facilitates the meta-analysis of different cohorts. When different chips are used for genotyping, imputation can equate the set of SNPs across studies. Results then can be combined together to increase the power of analysis by increasing the sample size. [23]

1.4 Statistical fine mapping

GWAS helps to identify a region on the genome that has the possibility of containing a causal variant. This is only the first step and additional statistical analysis should be performed to differentiate causal variants from the variants that are correlated with causal variants due to proximity. Most of the SNPs on microarray chips are variants that have a large linkage disequilibrium (LD) with the causal variants in their neighborhood. [24] LD can be defined as the together inheritance of the alleles within a haplotype more than it would be suggested by random chance. [25] Often LD patterns between SNPs are complex and it is not easy to identify the causal SNP. Statistical fine-mapping can help us unravel the causal variants. For performing fine mapping we need an association of region on the genome with a trait found in previous studies and the assumption that a casual variant exists. Firstly associations are discovered using GWAS. Hits over a threshold significance value (P-value $<$ generally $5 * 10^{-}8$) [26] are selected and the locations of these lead SNPs are tagged for fine mapping. A sample workflow for fine mapping is shown in the Figure 1.4 below.

Other than LD several other criteria can influence the power of fine mapping. Two main examples are SNP density and sample size. Sample size can be increased with a costly method of sampling more individuals or combining data from different studies at the cost of losing statistical power. For increasing SNP density we can use genotype imputation. An important factor while imputing genotype sit to choose an appropriate reference panel for the dataset at hand. Expectedly, the less accurate the imputation is, the less significant associations we will detect. [22]

Once significant SNPs are selected by fine-mapping we can proceed to decode their bi-

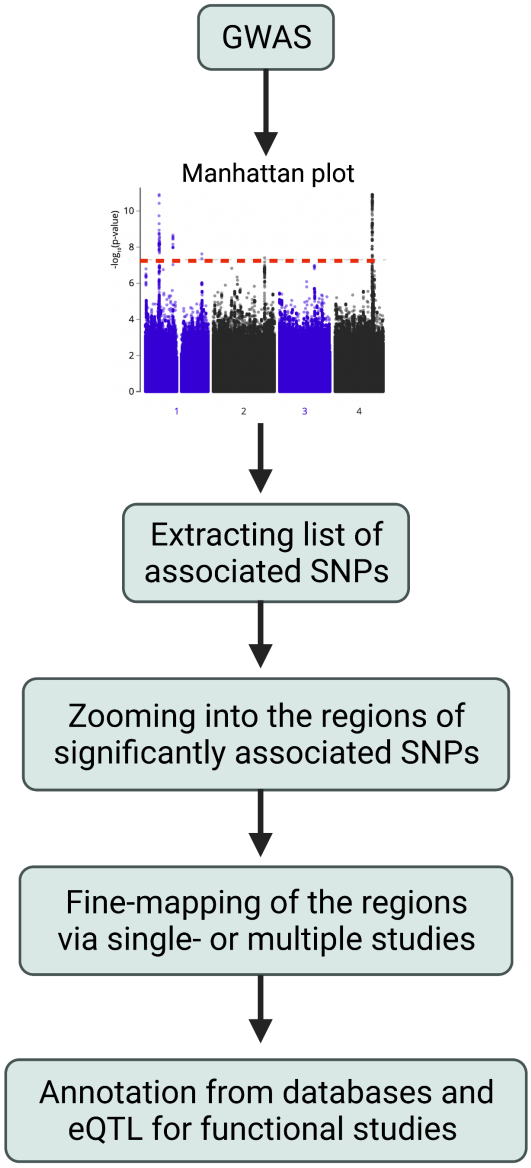

Figure 1.4: Fine mapping workflow. A representation of a genotype fine-mapping workflow which starts with GWAS analysis.

ological functions. Databases such as Gene Ontology [27] and ENCODE [28] can provide information about the enrichment of functional annotations. Generally, annotations are categorized into protein-coding and non-coding sequences. SNPs in coding regions can directly affect the conformation of the resulting protein while non-coding variation can have an effect on gene regulation. Non-coding annotations can be classified as, promoters, terminators, enhancers, transcription factor binding sites, epigenetic modification sites and etc.

1.5 X chromosome issues

Sexual dimorphism can be observed in many complex traits. Autoimmune disorders, cardiovascular diseases, and behavioral conditions demonstrate sex bias. Studies show that loci on the X chromosome have a higher probability of showing expression variance based on sex when compared to loci on autosomes. [29] Chromosome X is the 8 largest human chromosome with a length of 156 megabase pairs and 1669 discovered genes. [30] This accounts for around 5% total number of human genes. According to the Online Mendelian Inheritance in Man (OMIM) database, 7% of the diseases with known mechanisms are X-linked (Figure 1.5). [31]

Although the importance of the X chromosome in deciphering complex traits is well demonstrated by the above-mentioned facts, when it comes to analysis X chromosome is often excluded by researchers. The X chromosome has the least number of associations found in distinct loci in published GWAS studies after the Y chromosome. [4] Given its size and the number of genes it contains which is in turn similar to chromosome 7, it is very likely that we have yet to discover most of the associations on the X chromosome. Authors generally report a few common reasons for the exclusion of the X chromosome. Low coverage of chromosome X in microarray assays, lower genotyping accuracy compared to autosomes, and relatively challenging nature of analysis and interpretation of the data are some of them. Most of the new microarray assays now include a wide overage of the X chromosome. It is true that making sense out of the X chromosome data requires additional effort because of a higher rate of missing data, a higher frequency of chromosomal abnormalities, and dosage difference between male and female samples. Clustering algorithms generally perform worse on the X chromosome leading to lower accuracy scores. [4] Genotype imputation protocols for X chromosome are very similar to autosomes. The prime source of complexity comes from the dosage difference between males and females. Although most available imputation tools can in principle compensate for it are generally omitted from the software. In addition, the hemizygosity of males results in a decrease of the sample size by a quarter. [22] Combining all these reasons and the probability of obtaining enough publishable material only using autosomal data without spending extra effort, results in sex chromosomes lagging behind autosomes in research.

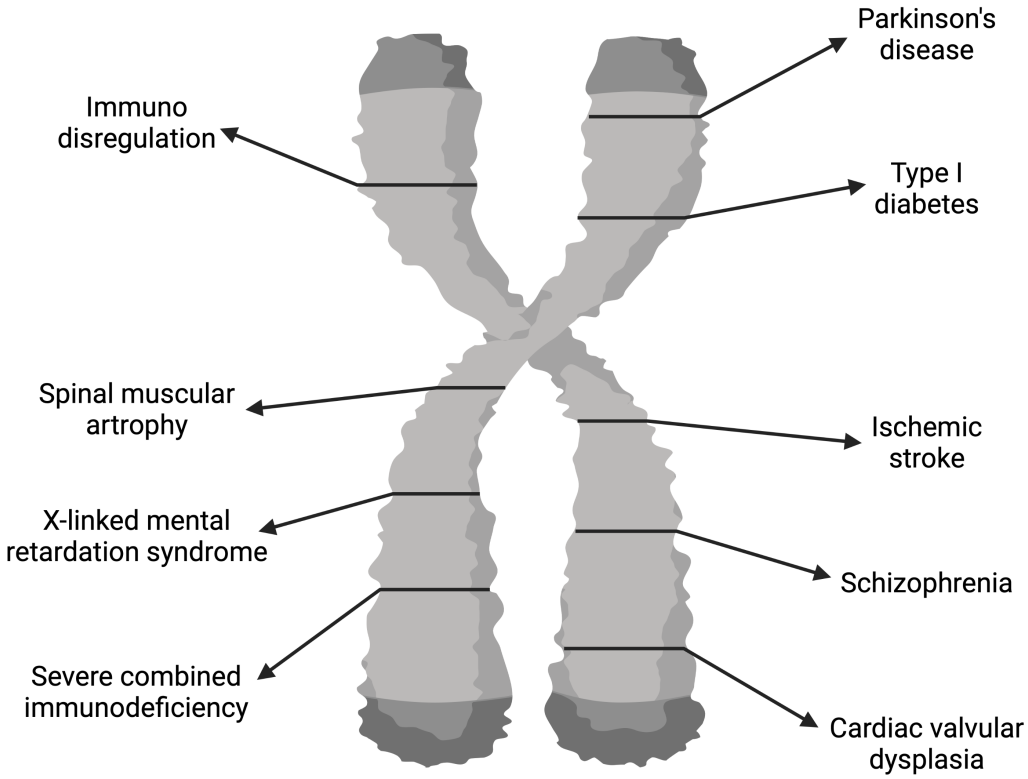

Figure 1.5: X chromosome associated complex traits.

2 Methods and result

We based the imputation and quality control stage of my project on the eQTL catalogue's gen-impute workflow developed by my supervisor Dr. Kaur Alasoo and added the ability to incorporate chromosome X to the workflow. [32] Figure 2.1 describes the summary of the work done in consecutive steps.

2.1 Preparing VCF files for eQTL analysis

Generally, genomics data is stored in Variant Call Format (VCF) format. VCF is a compressed tab-separated text file. It stores meta-information such as headers, sample IDs, and positions on the genome. Individual identification numbers are stored on rows while columns correspond to data such as the genotype of the individual at a given SNP. Figure 2.2 shows sample data stored in VCF format for genomic analysis. One of the most effective ways of working with VCF files is using BCFtools software. BCFtools is a freely available utility set that allows us to manipulate VCF files efficiently. [33]

Before performing statistical analysis on sequencing data, quality control measures need to be taken. For quality control steps we have used PLINK 1.9 software. Firstly we converted VCF files to plink files. We used the "–make-bed" flag of plink. The next step involved imputing the sex of the samples using PLINK. Following the separation of the X chromosome from the rest of the data we excluded pseudoautosomal regions - regions that are shared between X and Y chromosomes - from the next steps. We designated heterozygote haploid genotype to "missing" via PLINK. At this point, we extracted the list of female samples which will be used later on quality control steps. It is worth noting that we had to change the file names for the chromosome X multiple times during the process to ensure compatibility with the different tools we are using. Subsequently, we used the Genotype harmonizer command-line tool [35] to align our data with 1000 genomes project data which we used as reference. [36]

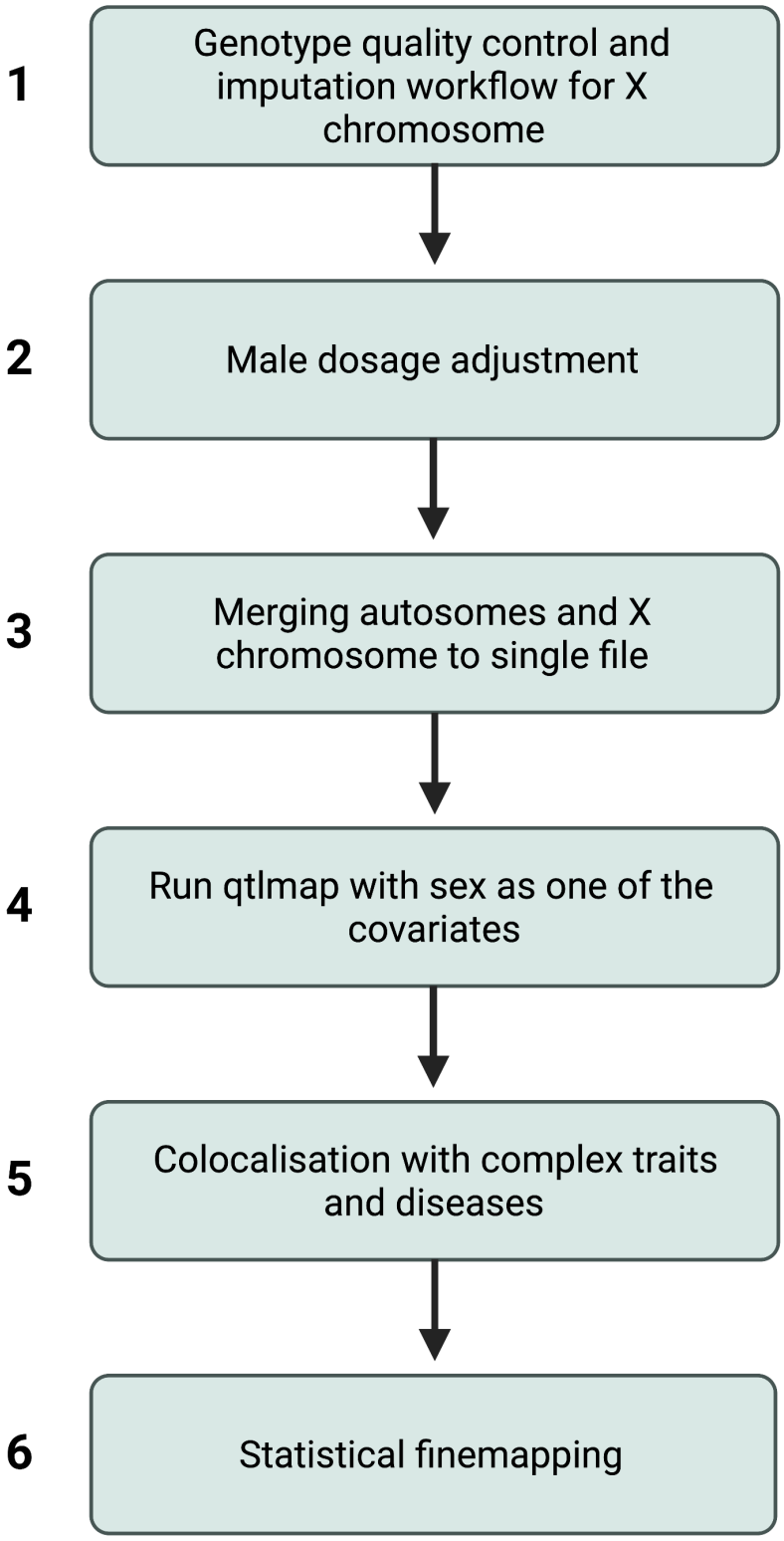

Figure 2.1: Summary of the project workflow.

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:
```

Figure 2.2: Sample genomic data stored in VCF format. [34]

Prior to proceeding to genotype imputation, we used the Eagle2 algorithm for haplotype phasing. [37] Haplotype phasing is the process of inferring haplotypes from given genotype data. Afterward, we used Minimac software for imputing genotypes in our project. [38] Minimac is a computationally efficient variant of the MaCH algorithm. It takes a VCF file as an input and outputs a VCF file with imputed genotypes. We used a reference panel from the 1000 Genomes project for our imputation. Following completing all the steps we used the merge function of BCFtools to firstly merge together dosage adjusted male samples to female samples and then X chromosome to the rest of the chromosomes. We filtered variants with missingness over 5% and minor allele frequency of less than 1 percent. We removed the coordinates of variants that passed quality control and used it to filer our data file to only contain quality control passed variants using BCFtools again.

I used VCF files from the GEUVADIS study [39] for preparing GEUVADIS data for eQTL analysis I first separated it into chromosomes. I used “bcftools view -r chr_” to split files. Since genomics files are very big in size, this step is necessary to avoid memory constraints and allow parallelization of the workflow. Despite having genotype data GEUVADIS files lacked dosage information for samples. We needed dosage information since qtlmap workflow used on the next steps depends on it to determine the number of alternative alleles in eQTL analysis. A given person can have a dosage of 0 if they lack the allele of interest, 1 if they are heterozygotes, and 2 if they are homozygotes. In spite of being very useful in manipulating VCF files, BCFtools does not have a straightforward way to calculate dosage based on genotype and add a dosage field

to files. I developed a script to tackle this problem and infer dosage from the genotype (Figure 2.3). This script altered the column named “GT” (genotype) to contain dosage information in the format of “GT:DS”. I used this script on all autosomes subsequently.

```
import pandas as pd
import numpy as np
import argparse

parser = argparse.ArgumentParser(description="Get DS from GT")

parser.add_argument("-i",type=str,help = "input file name")
parser.add_argument("-o",type=str,help = "output file name")
args = parser.parse_args()

data = pd.read_csv(str(args.i),sep='\t',comment="#",header =None)
data.iloc[:,8] = "GT:DS"

for n, d in data.iloc[:,9:].iteritems():

        new_list = []
        for i in d:
            i1 = i.split("|")
            ds = int(i1[0]) + int(i1[1])
            new_list.append(i+":"+str(ds))
        data = data.drop(n, axis=1)
        data[n] = new_list

data.to_csv(str(args.o), sep='\t',index = False)
```

Figure 2.3: Dosage from genotypes python script. This script calculates dosage from genotypes and add it to the GT column converting it to GT:DS format. The script is only applicable for autosomes and X chromosomes of females.

Before proceeding to the X chromosome I filtered VCF files based on sex to 2 different files. Although there are multiple ways that dosage difference between male and female can be addressed in our project we decided to multiply the dosage of male samples by a factor of 2. Having double the number of X chromsomescompared to males does not translate into having double the gene expression amount in females. During embryonic development, by the process called X inactivation, females transcriptionally disable one of their X chromsomes. Each copy is silenced in roughly the amount of cells. Subsequently females have mosaic of cells with different copy of X chromosome inactivated. This results in having the same dosage as males when it comes to loci found on X chromsome. Cite here Doubling male dosage in our workflow makes sure that the dosage balance is maintained throughout the analysis. [40] I developed a python script that can be used as a command-line tool to take a VCF file as an

input, manipulate the GT column which contains genotype information. The script does not alter genotype information while multiplying male dosage coefficients by 2. Similar to the autosomes it then adds a dosage field to the genotype column creating a column formatted as "GT:DS". Afterward, it outputs a VCF file with altered dosage information. Figure 2.4 shows the script I used to multiply male dosages.

```python
import pandas as pd
import numpy as np
import argparse

parser = argparse.ArgumentParser(description="Multiply male genotype dosage by 2")

parser.add_argument("-i",type=str,help = "input file name")
parser.add_argument("-o",type=str,help = "output file name")
args = parser.parse_args()

data = pd.read_csv(str(args.i),sep='\t',comment="#")

for n, d in data.iloc[:,9:].iteritems():

        new_list = []
        for i in d:
            print(i)
            i = i.split(":")
            i[1] = float(i[1])
            i[1] = str(i[1]*2)
            new_list.append(i)
        new_list = [':'.join(j) for j in new_list]
        data = data.drop(n, axis=1)
        data[n] = new_list

data.to_csv(str(args.o), sep='\t',index = False)
```

Figure 2.4: Dosage Multiplier python script. This script calculates dosage from genotypes and add it to the GT column converting it to GT:DS format. The script is only applicable for autosomes and X chromosomes of males.

Following the completion of dosage adjustment, I added headers back to the VCF files using "reheader" flag of BCFtools. I proceeded to reorder and index X chromosome files for males and females using "index" flag of BCFtools. I merged these files together as one before proceeding to create a single dosage adjusted file that included all chromosomes using the "merge" function.

2.2 QTLmap analysis

I submitted processed and QTLmap ready VCF files to Dr. Alasoo for eQTL fine-mapping analysis with sex as a covariate. I received a file that contains genetic variants that are likely to causally regulate gene expression. After filtering results that I received from him for the X

chromosome there are remaining 1884 eQTLs. They shared between 91 fine mapped credible sets. T perform colocalization analysis with complex traits and diseases I used two different GWAS datasets. FinnGen data can be accessed at https://r4.finngen.fi/top_hits and GWAS catalog data at https://www.ebi.ac.uk/gwas/docs/file-downloads. FinnGen dataset contained 668 associations on the X chromosome and the GWAS catalog had 1255 associations. The result of the colocalization analysts gave us 1 colocalization with FinnGen data and 7 with GWAS catalog. They are summarized in Table 2.1.

To further validate our eQTL signals we compared PLP2 hit to OpenTargets Genetic portal [41]. We compared our results to Phenome-wide association study of variant number 49171812 which take an SNP as an input and test its association against a large number of phenotypic variants. As seen from the PheWAS of PLP2 in Figure 2.5, the signal has been confirmed with multiple previous studies and is significantly associated with respiratory functions such as vital capacity, forced expiratory volume, and cardiovascular traits such as platelet distribution width.

Table 2.1: Colocalization of GEUVADIS eQTLS with FinnGen database and GWAS catalog. First 7 rows show the colocalization with GWAS catalog and the last row depicts colocalization with FinnGen data.

Variant Position	**GWAS trait**	**GWAS p-value**	**eQTL gene**	**eQTL pip value**
23773407	ribose-5-phosphate measurement, ribulose-5-phosphate measurement	5.000000e-07	ENSG00000233785	0.089159
49171812	platelet component distribution width	9.000000e-10	PLP2	0.500153
49187155	mean platelet volume	4.000000e-12	PLP2	0.500153
55528377	self reported educational attainment	1.000000e-09	KLF8	0.001232
55552777	mean corpuscular volume	3.000000e-08	KLF8	0.000281
119433739	osteitis deformans	1.000000e-07	SLC25A43	0.043446
155491696	factor VIII measurement	3.000000e-09	TMLHE	0.001712
53406445	E4_OBESITY_HYPER	8.818000e-08	HSD17B10	0.039949

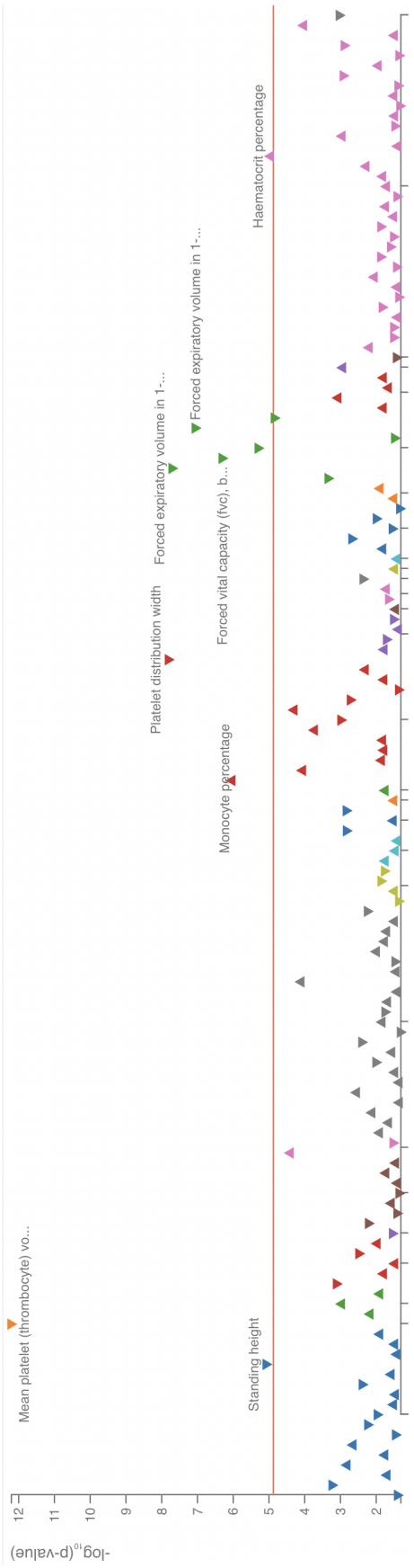

Figure 2.5: PLP2 PheWAS. Phenome-wide association study of variant number 49171812. Triangles over the red line (p-value $< 5\%$) represent significant associations

3 Discussion and conclusions

With the advance in the inclusion of more and more X chromosome SNPs in microarrays, we have access to large amounts of under-utilized data. By developing a robust genotype imputation workflow for the X chromosome we can take advantage of existing pipelines built for autosomes to analyze the X chromosome as well. In this project, we started to build this workflow and made significant advances towards its completion.

The single most important hurdle in utilizing X chromosome data is dosage dissimilarity between males and females. Existing software such as BCFtools does not provide a method to add and manipulate dosage data to VCF files. I have developed command-line tools to add dosage information to datasets while multiplying male dosage by two on the X chromosome to compensate for their heterozygosity. I used GEUVADIS data set for demonstrating the ability of the developed pipeline. The resulting dataset was. fully compatible with qtlmap workflow which is designed to work with autosomes.

After statistical fine-mapping, I demonstrated overlap between our eQTLs and two major databases in FinnGen and GWAS catalog. In total, we achieved 7 overlapping hits. We further validated the signal by cross-checking with the OpenTarget platform which summarises previous studies. Discovered associations were related to vital traits such as respiratory and cardiovascular functions demonstrating the potentially very useful information waiting to be discovered in underused X chromosome data.

It is worth noting that GEUVADIS data was collected just from human lymphoblastoid cells. [39] By utilizing different tissues and cell lines in eQTL catalogue [32] analysis can be expanded. It is likely that this will result in more significant associations discovered which then can be verified in laboratory settings. By identifying causal variants and decreasing the number of potential targets for drugs and therapies, computational workflows give an opportunity to focus research and funds to candidates with more likelihood of success

Theoretically, all aforementioned steps can be performed manually one after another. How-

ever, this would take an immense amount of time and organization. It is very easy to make a mistake while dealing with many similar file names and different formats. This kind of approach would also take a lot of time since you have to run every single step one by one after each other. For reproducibility of the research and running pipelines in parallel in high-performance computing clusters (HPC), we used Nextflow workflow manager. [42] We integrated our scripts into one modular pipeline. Not only it allows parts of our project to be used separately and modified to fit the needs of researchers but it also makes our project very easily reproducible and faster to run. This step was not fully completed due to time constraints and further effort is needed to integrate pipelines within nextflow environment.

Acknowledgement

I would like to extend my gratitude to my supervisor Kaur Alasoo. I've never learned as much from anyone as I learned from him about computational biology in the past months. I would like to thank professor Jaak Vilo for introducing me to Kaur, and professor Gholamreza Anbarjafari for always advising and supporting me. Of course, none of these would be possible without encouragement from my dear family. I'd like to also thank Artemi Maljavin for his continuous efforts in this process, and Hanno Evard for being a great mentor. My doctors, who did their very best to make me physically ready in time for our defense, I salute you. Finally, I'm forever grateful for my program director, professor Ilona Faustova. Without her guidance in the past 3 years and aiding me whenever I needed it the most I would have not made it through.

References

[1] J. L. Badano and N. Katsanis, "Beyond mendel: An evolving view of human genetic disease transmission", *Nature Reviews Genetics*, vol. 3, no. 10, pp. 779–789, 2002.

[2] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, "Human genetic variation and its contribution to complex traits", *Nature Reviews Genetics*, vol. 10, no. 4, pp. 241–251, 2009.

[3] K. J. Mitchell, "What is complex about complex disorders?", *Genome biology*, vol. 13, no. 1, pp. 1–11, 2012.

[4] A. L. Wise, L. Gyi, and T. A. Manolio, "Exclusion: Toward integrating the x chromosome in genome-wide association analyses", *The American Journal of Human Genetics*, vol. 92, no. 5, pp. 643–647, 2013.

[5] A. O. Edwards, R. Ritter, K. J. Abel, A. Manning, C. Panhuysen, and L. A. Farrer, "Complement factor h polymorphism and age-related macular degeneration", *Science*, vol. 308, no. 5720, pp. 421–424, 2005.

[6] M. C. Mills and C. Rahal, "A scientometric review of genome-wide association studies", *Communications biology*, vol. 2, no. 1, pp. 1–11, 2019.

[7] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies", *Nature Reviews Genetics*, vol. 20, no. 8, pp. 467–484, 2019.

[8] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases", *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[9] A. Mahajan, J. Wessel, S. M. Willems, W. Zhao, N. R. Robertson, A. Y. Chu, W. Gan, H. Kitajima, D. Taliun, N. W. Rayner, *et al.*, "Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes", *Nature genetics*, vol. 50, no. 4, pp. 559–571, 2018.

[10] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox, "Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from gwas", *PLoS Genet*, vol. 6, no. 4, e1000888, 2010.

[11] Y. Liu, X. Liu, Z. Zheng, T. Ma, Y. Liu, H. Long, H. Cheng, M. Fang, J. Gong, X. Li, *et al.*, "Genome-wide analysis of expression qtl (eqtl) and allele-specific expression (ase) in pig muscle identifies candidate genes for meat quality traits", *Genetics Selection Evolution*, vol. 52, no. 1, pp. 1–11, 2020.

[12] N. Shan, Z. Wang, and L. Hou, "Identification of trans-eqtls using mediation analysis with multiple mediators", *BMC bioinformatics*, vol. 20, no. 3, pp. 87–97, 2019.

[13] M. Kellis, *Lecture 15 – Mediation, eQTLs population genetics, history*, Machine Learning in Genomics lecture slides, MIT, Fall 2020.

[14] J. T. Burdick, W.-M. Chen, G. R. Abecasis, and V. G. Cheung, "In silico method for inferring genotypes in pedigrees", *Nature genetics*, vol. 38, no. 9, pp. 1002–1004, 2006.

[15] Y. Li, C. Willer, S. Sanna, and G. Abecasis, "Genotype imputation", *Annual review of genomics and human genetics*, vol. 10, pp. 387–406, 2009.

[16] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes", *Nature genetics*, vol. 39, no. 7, pp. 906–913, 2007.

[17] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "Mach: Using sequence and genotype data to estimate haplotypes and unobserved genotypes", *Genetic epidemiology*, vol. 34, no. 8, pp. 816–834, 2010.

[18] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland, "Score tests for association between traits and haplotypes when linkage phase is ambiguous", *The American Journal of Human Genetics*, vol. 70, no. 2, pp. 425–434, 2002.

[19] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation plink: Rising to the challenge of larger and richer datasets", *Gigascience*, vol. 4, no. 1, s13742–015, 2015.

[20] D. L. Nicolae, "Testing untyped alleles (tuna)—applications to genome-wide association studies", *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, vol. 30, no. 8, pp. 718–727, 2006.

[21] S. R. Browning, "Multilocus association mapping using variable-length markov chains", *The American Journal of Human Genetics*, vol. 78, no. 6, pp. 903–913, 2006.

[22] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies", *Nature Reviews Genetics*, vol. 11, no. 7, pp. 499–511, 2010.

[23] E. Zeggini, L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen, *et al.*, "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes", *Nature genetics*, vol. 40, no. 5, pp. 638–645, 2008.

[24] D. MacArthur, T. Manolio, D. Dimmock, H. Rehm, J. Shendure, G. Abecasis, D. Adams, R. Altman, S. Antonarakis, E. Ashley, *et al.*, "Guidelines for investigating causality of sequence variants in human disease", *Nature*, vol. 508, no. 7497, pp. 469–476, 2014.

[25] D. J. Schaid, W. Chen, and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping", *Nature Reviews Genetics*, vol. 19, no. 8, pp. 491–504, 2018.

[26] T. Zeller, S. Blankenberg, and P. Diemert, "Genomewide association studies in cardiovascular disease—an update 2011", *Clinical chemistry*, vol. 58, no. 1, pp. 92–103, 2012.

[27] G. O. Consortium, "The gene ontology: Enhancements for 2011", *Nucleic acids research*, vol. 40, no. D1, pp. D559–D564, 2012.

[28] Y. Zhao, E. Schaafsma, and C. Cheng, "Applications of encode data to systematic analyses via data integration", *Current opinion in systems biology*, vol. 11, pp. 57–64, 2018.

[29] K. R. Kukurba, P. Parsana, B. Balliu, K. S. Smith, Z. Zappala, D. A. Knowles, M.-J. Favé, J. R. Davis, X. Li, X. Zhu, *et al.*, "Impact of the x chromosome and sex on regulatory variation", *Genome research*, vol. 26, no. 6, pp. 768–777, 2016.

[30] K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, *et al.*, "Telomere-to-telomere assembly of a complete human x chromosome", *Nature*, vol. 585, no. 7823, pp. 79–84, 2020.

[31] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders", *Nucleic acids research*, vol. 43, no. D1, pp. D789–D798, 2015.

[32] N. Kerimov, J. D. Hayhurst, J. R. Manning, P. Walter, L. Kolberg, K. Peikova, M. Samoviča, T. Burdett, S. Jupp, H. Parkinson, *et al.*, "Eqtl catalogue: A compendium of uniformly processed human gene expression and splicing qtls", *BioRxiv*, 2020.

[33] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools", *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[34] H. M. Elshazly, "Optimizing bioinformatics variant analysis pipeline for clinical use", M.S. thesis, Nile University, 2016.

[35] P. Deelen, M. J. Bonder, K. J. van der Velde, H.-J. Westra, E. Winder, D. Hendriksen, L. Franke, and M. A. Swertz, "Genotype harmonizer: Automatic strand alignment and format conversion for genotype data integration", *BMC research notes*, vol. 7, no. 1, pp. 1–4, 2014.

[36] 1. G. P. Consortium *et al.*, "A global reference for human genetic variation", *Nature*, vol. 526, no. 7571, p. 68, 2015.

[37] P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A. Reshef, H. K. Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, *et al.*, "Reference-based phasing using the haplotype reference consortium panel", *Nature genetics*, vol. 48, no. 11, p. 1443, 2016.

[38] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing", *Nature genetics*, vol. 44, no. 8, pp. 955–959, 2012.

[39] T. Lappalainen, M. Sammeth, M. R. Friedländer, P. Ac't Hoen, J. Monlong, M. A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, *et al.*, "Transcriptome and genome sequencing uncovers functional variation in humans", *Nature*, vol. 501, no. 7468, pp. 506–511, 2013.

[40] H. Fang, C. M. Disteche, and J. B. Berletch, "X inactivation and escape: Epigenetic and structural features", *Frontiers in cell and developmental biology*, vol. 7, p. 219, 2019.

[41] D. Ochoa, A. Hercules, M. Carmona, D. Suveges, A. Gonzalez-Uriarte, C. Malangone, A. Miranda, L. Fumis, D. Carvalho-Silva, M. Spitzer, *et al.*, "Open targets platform: Supporting systematic drug–target identification and prioritisation", *Nucleic Acids Research*, vol. 49, no. D1, pp. D1302–D1310, 2021.

[42] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows", *Nature biotechnology*, vol. 35, no. 4, pp. 316–319, 2017.

Non-Exclusive licence to reproduce thesis and make thesis public

I, **Nihat Aliyev**,

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,

   Developing a computational workflow for eQTL analysis on the X chromosome

   supervised by PhD 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.

Nihat Aliyev
Tartu, 20.05.2021