DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
47

NURLAN KERIMOV

Building a catalogue of molecular quantitative trait loci to interpret complex trait associations

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

47

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
47

NURLAN KERIMOV

Building a catalogue of molecular quantitative trait loci to interpret complex trait associations

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in informatics on 07 October, 2023 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor:

Dr. Kaur Alasoo
University of Tartu, Estonia

Opponents:

Prof. Gregory C. Gibson
Georgia Institute of Technology, United States of America

Dr. Emma Davenport
Wellcome Sanger Institute, United Kingdom

The public defence will take place on 13 November 2023 at 14:00 in University of Tartu DELTA house room number 1005.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print) ISSN 2806-2345 (PDF)
ISBN 978-9916-27-382-1 (print) ISBN 978-9916-27-383-8 (PDF)

University of Tartu Press
http://www.tyk.ee/

If you flip 1,000 fair coins 10 times each, statistically, one of them is likely to land on heads in all ten flips. This coin will then tour conferences, giving talks about the dedication and effort it put into achieving such notable success

Anonymous

ABSTRACT

In the pursuit of understanding the biological mechanisms underlying complex diseases, this study presents a novel approach involving the use of quantitative trait loci (QTL) analysis to enhance the interpretation of findings from Genome-Wide Association Studies (GWAS). Recognising the interpretational challenges posed by the numerous genetic variations with small individual effects identified by GWAS, we developed a compendium of uniformly processed human molecular QTLs. Our eQTL Catalogue comprises data from 74 distinct cell types and tissues and several environmental stimuli, with all 127 datasets processed uniformly for optimal downstream analyses. Notably, the construction of this resource was enabled by our development of robust, open-source scientific pipelines, which play a crucial role not only in the generation of the eQTL Catalogue but also in facilitating the adoption of pipeline reusability and federated analysis approaches. In response to feedback on difficulties in QTL signal interpretation, we also generated QTL coverage plots for all independent genetic signals and related molecular traits. The eQTL Catalogue has proven invaluable in various research projects, aiding in the interpretation of GWAS findings and contributing to the understanding of the genetic basis of complex traits. Furthermore, the infrastructure supports quick re-analysis as new methods emerge, demonstrating the Catalogue's flexibility and adaptability in genomic research.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AMDHD1	Amidohydrolase Domain Containing 1 (gene)
API	Application Programming Interface
BAM	Binary Alignment/Map (file format)
CAGE	Cap Analysis of Gene Expression
CI	Continuous Integration
CPU	Central Processing Unit
CYP2R1	Cytochrome P450 Family 2 Subfamily R Member 1 (gene)
DNA	Deoxyribonucleic Acid
DSL	Digital Subscriber Line
FDR	False Discovery Rate
FinnGen	A large public-private genome research project in Finland
FTP	File Transfer Protocol
GUI	Graphical User Interface
GWAS	Genome-Wide Association Study
HAL	Histidine Ammonia-Lyase (gene)
HCP	Hidden Covariates with Prior
HPC	High-Performance Computing
IGV	Integrative Genomics Viewer
INT	Inverse Normal Transformation
IO	Input/Output
iPSCs	Induced Pluripotent Stem Cells
LBF	Log Bayes Factor
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MANE	Matched Annotation from NCBI and EMBL-EBI
MBV	Match BAM to VCF
MDS	Multidimensional Scaling
NMD	Nonsense-Mediated mRNA Decay
PCA	Principal Component Analysis
PEER	Probabilistic Estimation of Expression Residuals
PIP	Posterior Inclusion Probability
PP	Posterior Probability
QC	Quality Control
QTL	Quantitative Trait Locus
REST	Representational State Transfer
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
SNP	Single Nucleotide Polymorphism
SuSiE	Sum of Single Effects
SVA	Surrogate Variable Analysis
TF	Transcription Factor
TPM	Transcripts per Million
UK	United Kingdom
USD	United States Dollar
VCF	Variant Call Format
WGS	Whole Genome Sequencing

LIST OF PUBLICATIONS INCLUDED IN THE THESIS

Publications included in the thesis

I. **Nurlan Kerimov**, James D. Hayhurst, Kateryna Peikova, Jonathan R. Manning, Peter Walter, Liis Kolberg, Marija Samoviča, Manoj Pandian Sakthivel, Ivan Kuzmin, Stephen J. Trevanion, Tony Burdett, Simon Jupp, Helen Parkinson, Irene Papatheodorou, Andrew D. Yates, Daniel R. Zerbino, and Kaur Alasoo. 2021. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* 53, 9 (September 2021), 1290–1299.

II. **Nurlan Kerimov**, Ralf Tambets, James D. Hayhurst, Ida Rahu, Peep Kolberg, Uku Raudvere, Ivan Kuzmin, Anshika Chowdhary, Andreas Vija, Hans J. Teras, Masahiro Kanai, Jacob Ulirsch, Mina Ryten, John Hardy, Sebastian Guelfi, Daniah Trabzuni, Sarah Kim-Hellmuth, Will Rayner, Hilary Finucane, Hedi Peterson, Abayomi Mosaku, Helen Parkinson, and Kaur Alasoo. eQTL Catalogue 2023: New datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs, *PLoS Genet*. 19 (2023) e1010932.

III. Mitja I. Kurki, Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati M. Donner, Mary P. Reeve, Hannele Laivuori, Mervi Aavikko, Mari A. Kaunisto, Anu Loukola, Elisa Lahtela, Hannele Mattsson, Päivi Laiho, Pietro Della Briotta Parolo, Arto A. Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O. Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, Wei Zhou, Sanni Ruotsalainen, Kalle Pärn, Tero Hiekkalinna, Sami Koskelainen, Teemu Paajanen, Vincent Llorens, Javier Gracia-Tabuenca, Harri Siirtola, Kadri Reis, Abdelrahman G. Elnahas, Benjamin Sun, Christopher N. Foley, Katriina Aalto-Setälä, Kaur Alasoo, Mikko Arvas, Kirsi Auro, Shameek Biswas, Argyro Bizaki-Vallaskangas, Olli Carpen, Chia-Yen Chen, Oluwaseun A. Dada, Zhihao Ding, Margaret G. Ehm, Kari Eklund, Martti Färkkilä, Hilary Finucane, Andrea Ganna, Awaisa Ghazal, Robert R. Graham, Eric M. Green, Antti Hakanen, Marco Hautalahti, Åsa K. Hedman, Mikko Hiltunen, Reetta Hinttala, Iiris Hovatta, Xinli Hu, Adriana Huertas-Vazquez, Laura Huilaja, Julie Hunkapiller, Howard Jacob, Jan-Nygaard Jensen, Heikki Joensuu, Sally John, Valtteri Julkunen, Marc Jung, Juhani Junttila, Kai Kaarniranta, Mika Kähönen, Risto Kajanne, Lila Kallio, Reetta Kälviäinen, Jaakko Kaprio, FinnGen, **Nurlan Kerimov**, Johannes Kettunen, Elina Kilpeläinen, Terhi Kilpi, Katherine Klinger, Veli-Matti Kosma, Teijo Kuopio, Venla Kurra, Triin Laisk, Jari Laukkanen, Nathan Lawless, Aoxing Liu, Simonne Longerich, Reedik Mägi, Johanna Mäkelä, Antti Mäkitie, Anders Malarstig, Arto Mannermaa, Joseph Maranville, Athena Matakidou, Tuomo Meretoja, Sahar V. Mozaffari, Mari E. K. Niemi, Marianna Niemi, Teemu Niiranen,

Christopher J. O Donnell, Ma En Obeidat, George Okafo, Hanna M. Ollila, Antti Palomäki, Tuula Palotie, Jukka Partanen, Dirk S. Paul, Margit Pelkonen, Rion K. Pendergrass, Slavé Petrovski, Anne Pitkäranta, Adam Platt, David Pulford, Eero Punkka, Pirkko Pussinen, Neha Raghavan, Fedik Rahimov, Deepak Rajpal, Nicole A. Renaud, Bridget Riley-Gillis, Rodosthenis Rodosthenous, Elmo Saarentaus, Aino Salminen, Eveliina Salminen, Veikko Salomaa, Johanna Schleutker, Raisa Serpi, Huei-Yi Shen, Richard Siegel, Kaisa Silander, Sanna Siltanen, Sirpa Soini, Hilkka Soininen, Jae Hoon Sul, Ioanna Tachmazidou, Kaisa Tasanen, Pentti Tienari, Sanna Toppila-Salmi, Taru Tukiainen, Tiinamaija Tuomi, Joni A. Turunen, Jacob C. Ulirsch, Felix Vaura, Petri Virolainen, Jeffrey Waring, Dawn Waterworth, Robert Yang, Mari Nelis, Anu Reigo, Andres Metspalu, Lili Milani, Tõnu Esko, Caroline Fox, Aki S. Havulinna, Markus Perola, Samuli Ripatti, Anu Jalanko, Tarja Laitinen, Tomi P. Mäkelä, Robert Plenge, Mark McCarthy, Heiko Runz, Mark J. Daly, and Aarno Palotie. 2023. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* 613, 7944 (January 2023), 508–518.

IV. Masahiro Kanai, Jacob C. Ulirsch, Juha Karjalainen, Mitja Kurki, Konrad J. Karczewski, Eric Fauman, Qingbo S. Wang, Hannah Jacobs, François Aguet, Kristin G. Ardlie, **Nurlan Kerimov**, Kaur Alasoo, Christian Benner, Kazuyoshi Ishigaki, Saori Sakaue, Steven Reilly, Yoichiro Kamatani, Koichi Matsuda, Aarno Palotie, Benjamin M. Neale, Ryan Tewhey, Pardis C. Sabeti, Yukinori Okada, Mark J. Daly, Hilary K. Finucane, The BioBank Japan Project, and FinnGen. 2021. Insights from complex trait fine-mapping across diverse populations. *bioRxiv,* September 2021

My contributions to these publications

Publication I – In the creation of the eQTL Catalogue, which consolidates uniformly processed human gene expression and splicing quantitative trait loci, I served as one of the principal developers. I was responsible for the design and development of the scientific pipelines required for processing the datasets contained in this resource. My duties also included meticulous quality control of the majority of the studies and the assurance of process repeatability for future resource custodians. I concluded my contribution by conducting the downstream analysis of these datasets and co-wrote the manuscript that reported the findings with K. Alasoo.

Publication II – I was responsible for enhancing most of the QTL analysis pipelines to produce the necessary data for QTL visualisations. I designed and constructed an additional pipeline for systematically generating visualisations (coverage plots), successfully addressing all technical issues related to the pipeline's scalability. I also furnished the scripts illustrating how the data generated by this visualisation pipeline can be repurposed, thereby providing future users with a solid foundation for creating dynamic QTL visualisations. Lastly, I created all the figures for the manuscript and co-wrote the manuscript with K. Alasoo.

Publication III and IV – For these two publications, my involvement revolved around the integration of SuSiE fine mapping into our QTL analysis pipeline. This included reanalysing all the quality-controlled eQTL Catalogue datasets using the updated pipelines and delivering fine mapping results for these projects.

Publications not included in the thesis

I. Liis Kolberg, **Nurlan Kerimov**, Hedi Peterson, and Kaur Alasoo. 2020. Co-expression analysis reveals interpretable gene modules controlled by trans-acting genetic variants. *Elife* 9, (September 2020). DOI:https://doi.org/10.7554/eLife.58705

II. Nightingale Health Biobank Collaborative Group, Jeffrey C. Barrett, Tõnu Esko, Krista Fischer, Luke Jostins-Dean, Pekka Jousilahti, Heli Julkunen, Tuija Jääskeläinen, **Nurlan Kerimov**, Sini Kerminen, Anastassia Kolde, Harri Koskela, Jaanika Kronberg, Sara N. Lundgren, Annamari Lundqvist, Valtteri Mäkelä, Kristian Nybo, Markus Perola, Veikko Salomaa, Kirsten Schut, Maiju Soikkeli, Pasi Soininen, Mika Tiainen, Taavi Tillmann, Peter Würtz, and the Estonian Biobank Research Team. 2023. Metabolomic and genomic prediction of common diseases in 477,706 participants in three national biobanks. *medRxiv*. June 2023

INTRODUCTION

The holy grail of genomics research is to find the biological mechanisms that explain complex diseases so that they can be prevented or treated. Researchers have approached this extremely complex problem from several angles. One of the most widely-used and relatively successful approaches is GWAS. GWAS identifies genetic variants associated with complex traits (i.e. GWAS hits), such as diseases. However, there are usually many genetic variants associated with a trait, and these associations have small effects individually, which is hard to interpret. Hence, additional methods are necessary to explain these GWAS hits. One method is to find associations of genetic variants with expression levels of genes (i.e. QTL analysis) and compare these associations with GWAS hits (i.e. colocalisation analysis), under the assumption that the genetic variant affects the complex trait by regulating gene expression first. Simply put, genetic variants regulate expression of a gene, which results in the production of a certain protein, which in turn affects the complex trait. However, since the regulation of the genes happens differently in different biological contexts (i.e. cell types, tissues or conditions), if the colocalisation analysis is performed in the "wrong" context, it can be impossible to find the desired colocalisation signal.

The main goal of this thesis is to share the best methods and lessons learned from building scientific procedures to tackle the problems faced in QTL analysis. Due to the complexity of QTL analysis, which requires a deep understanding of the subject, the thesis will first provide comprehensive background knowledge in the initial three chapters. This will be followed by an in-depth exploration of the technical aspects involved in developing scientific pipelines for conducting QTL analysis on vast amounts of data.

In this dissertation, I will initially provide a brief overview of the fundamentals of molecular biology (Chapter 1), then proceed to describe the approaches and techniques used in order to interpret GWAS signals. This will encompass the differences between GWAS and QTL studies, colocalisation and fine-mapping techniques and will outline the key challenges encountered when utilising these approaches (Chapter 2). In Chapter 3, I will thoroughly explore the QTL analysis procedure, beginning with an overview of the entire workflow and its three main components: quantification, quality control and normalisation, and QTL mapping. The quantification section will discuss the workflow inputs, their categorisation as molecular traits and the methods used to transform RNA sequencing (RNA-seq) data into numerical tabular data. This will be followed by an examination of quality control and normalisation. Finally, I will delve into the intricacies of the QTL mapping process and subsequent post-processing stages.

The creation of the eQTL Catalogue resource would have not been possible without the modern infrastructure of scientific pipelines. In Chapter 4, I will start by defining these infrastructures and tracing the evolution of pipelines, followed by a comprehensive discussion of the properties of computational

infrastructures and scientific pipelines. Furthermore, I will address the latest advancements in cloud computing and containerisation technologies, which have significantly contributed to the efficient execution and scalability of these pipelines. The chapter will conclude with an analysis of the current challenges in pipeline development and potential strategies for avoiding or overcoming these obstacles.

In Chapter 5, I will provide a concise overview of the key aspects of Publication I and elaborate on noteworthy points that did not make it into the published work. These primarily involve the numerous decisions made during the development of the eQTL Catalogue workflow and the subsequent impact on the pipeline properties discussed in Chapter 4. Additionally, a subchapter has been included to shed light on the often-overlooked efforts and lessons learned throughout the pipeline development process. Similarly, Chapter 6 will discuss enhancing the interpretability of QTLs through visualisation, as detailed in Publication II. Following a brief introduction to the problem we aim to address, I will describe the visualisation, its construction and the primary challenges encountered during its development that were not included in the published work.

Chapter 7 will highlight the impact of the eQTL Catalogue as a valuable resource for the scientific community. In this chapter, I will delve into its various applications in diverse scientific studies and how other researchers have utilised the catalogue. Additionally, I will emphasise the significance of having a robust infrastructure in place to facilitate the rapid execution of pipelines for new experiments. The dissertation will conclude with a discussion chapter, in which I will address the essential aspects of our work, its limitations and prospective future developments for this valuable resource.

Standing on the shoulders of giants

Bernard of Chartres

1. BACKGROUND

The field of genetics has made significant progress in understanding how traits are inherited and how they vary between individuals. At the core of this progress lies the Central Dogma of molecular biology. A fundamental understanding of gene structure and genetic variation is also essential for grasping the impact of genetics on the variation of complex traits. In this chapter, we will provide a brief introduction to the Central Dogma, followed by an exploration of gene structure and genetic variation.

1.1. Central dogma of molecular biology

The central dogma of molecular biology is a fundamental principle that describes the flow of genetic information within cells. It consists of two main processes: RNA transcription and protein translation. Transcription is the process of creating RNA from a DNA segment, and translation is the process of creating a protein sequence from a transcribed RNA template. To summarise, the central dogma states that genetic information flows unidirectionally from DNA to RNA to protein (Figure 1) [1]. Although there are exceptions to this rule, such as RNA-dependent RNA synthesis and reverse transcription, the central dogma remains the cornerstone of molecular biology [2,3]. The importance of the central dogma is highlighted by the fact that it underpins the entire field of molecular biology and is essential for understanding the mechanisms of gene expression and regulation [4,5].

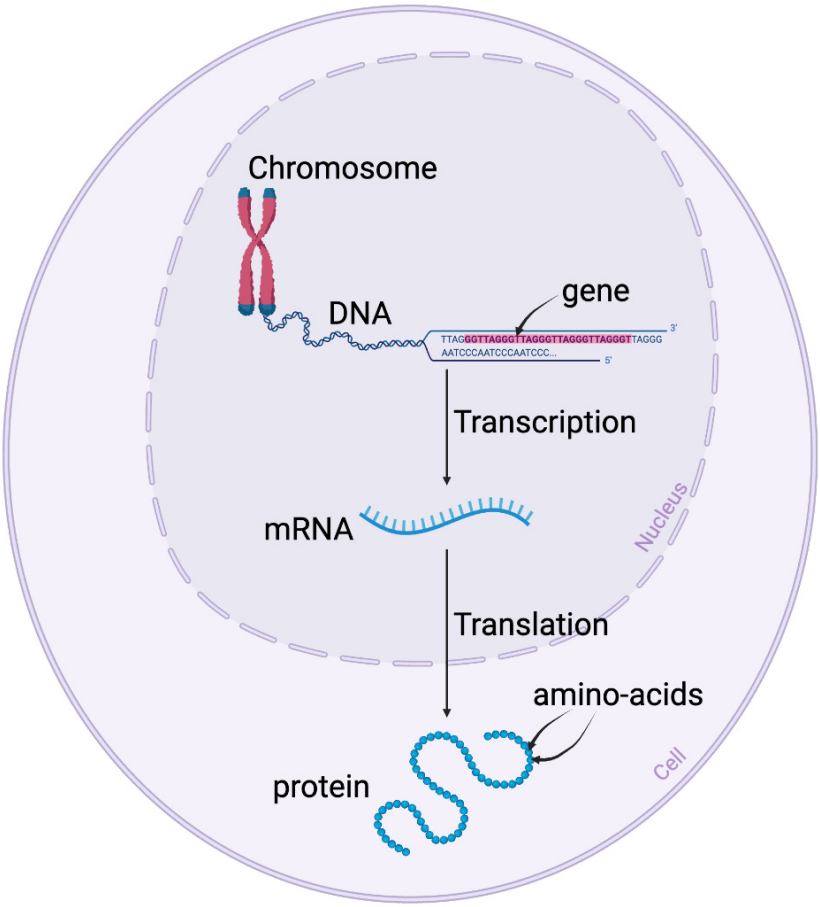

Figure 1: Central Dogma of Molecular Biology: A DNA sequence called a gene is transcribed into mRNA. Then, the mRNA molecule leaves the nucleus of the cell and is translated into amino acids, which chain together to form a protein. The figure is created using the BioRender application (BioRender.com, 2023)

1.2. Introduction to genetic variation

The fundamental processes of life are encoded within genetic material known as the genome, which is composed of deoxyribonucleic acid (DNA) molecules. DNA consists of two complementary nucleotide strands, forming a double helix structure, with nucleotides containing one of four distinct nucleobases: adenine (A), thymine (T), guanine (G), and cytosine (C). These nucleobases, often simply referred to as bases, serve as primary constituents of DNA and ensure the complementary nature of the strands. Human cells are diploid, meaning that each chromosome has a duplicate that originates in either parent. In total, there are 23 pairs of chromosomes in every human cell: 22 pairs of autosomes and one pair of sex chromosomes. Because of the diploid nature, every genetic variation in a specific human genome can appear zero, one or two times, and

this is referred to as that individual's genotype. A single set of 23 human chromosomes encompasses approximately 3.2 billion AT and GC base pairs connected in sequences [6–8]. Decoding this information is essential for understanding the mechanisms of life and represents a significant challenge in human health research.

Genetic variation refers to the differences in DNA sequences between individuals or populations [9]. Any two individuals are known to have ~99.5% identical DNA, whereas any human shares ~98.7% of the DNA with a chimpanzee [10]. Although these differences in proportion may seem small, in absolute numbers, there are more than 100 million genetic variants in humans, effectively resulting in an infinite number of allele combinations. These genetic variants mainly consist of three types: single-nucleotide polymorphisms (SNPs), insertion-deletions polymorphisms (INDELs), and structural variants (SVs). A typical human genome differs from the reference human genome at 4.1–5.0 million sites and more than 99.9% of this variation consists of SNPs and short INDELs. The remaining <0.01% consists mainly of 2,100 to 2,500 SVs, which in total affects more bases than the SNPs, around 20 million bases of sequence [11].

The genetic makeup of an individual plays a crucial role in determining their physical and physiological characteristics as well as their likelihood of developing certain diseases [12–14].

1.2.1. Linkage disequilibrium

Linkage disequilibrium (LD) is a term used in genetics to describe the non-random association between alleles at different loci (i.e. in different regions in the DNA) in a population. The term “linkage” refers to the physical proximity of the loci on the same chromosome, while “disequilibrium” refers to the non-random association of alleles at these loci (Figure 2). The reason for this is that the loci which are physically close together on the chromosome make it more likely that they will be inherited together as a unit, rather than being shuffled independently during meiosis[1]. As a result, certain combinations of alleles may be overrepresented or underrepresented in the population compared to what would be expected if the loci were independent.

LD also has non-negligible implications for genetic association studies, as it can impact the power and accuracy of these studies. Specifically, LD makes it challenging to pinpoint the exact causal variant(s), as the association may be driven by a nearby causal variant that is in LD with the tested variant. This phenomenon is known as “genetic confounding” and can be partially addressed using statistical methods such as conditional analysis or fine-mapping, which aim to identify the true causal variant(s) underlying an association signal, although many technical challenges remain (Publication IV) [15,16].

[1] Meiosis: a type of cell division that results in four genetically diverse haploid daughter cells, each with half the number of chromosomes as the parent cell, which are used as gametes in sexual reproduction.

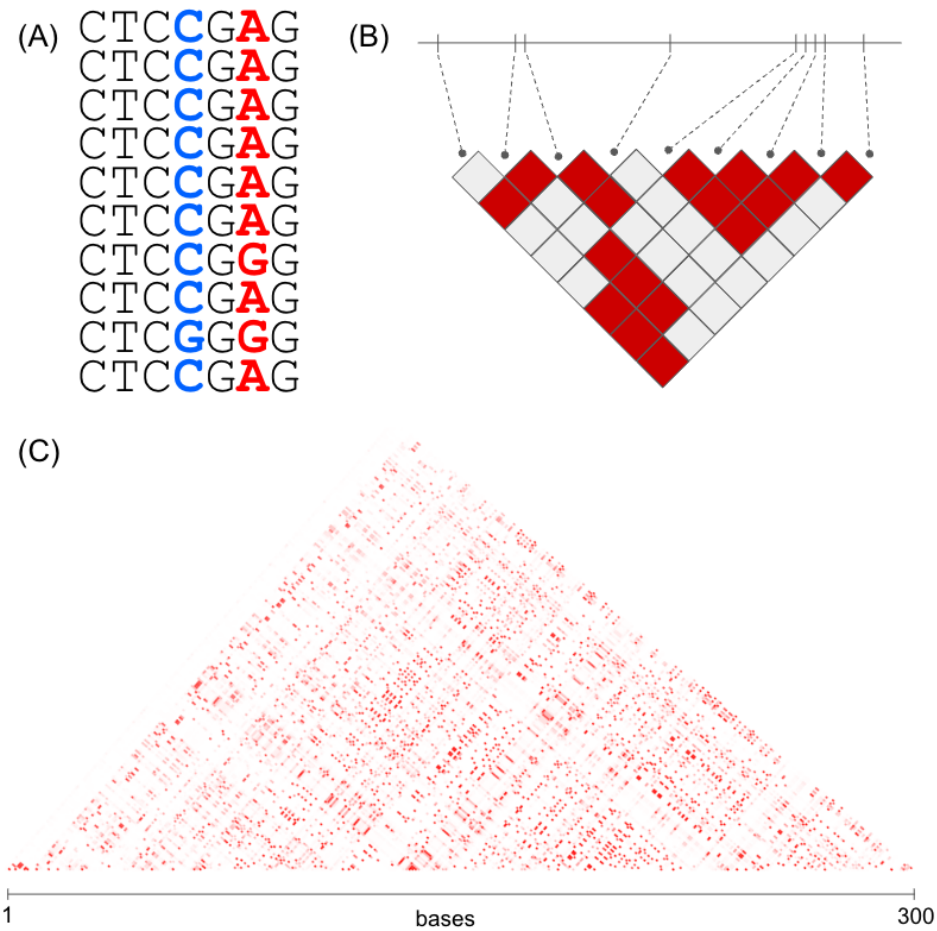

Figure 2: Example of LD and its effect on population structure. (A) Fragments of 10 samples from a population. Only seven bases are shown in the example and there are only two sites where these fragments are different (i.e. genetic variants), in positions 4 and 6 (in blue and red colours respectively). It can be observed that the "C" allele in position 4 and the "A" allele in position 6 are linked: when we observe the "C" allele in position 4 we usually observe the "A" allele in position 6; however, when we observe the "A" allele in position 6, we **always** observe "C" allele in position 4. (B) A simulated representation of LD between nine variants in a region. Dark red squares represent high correlation between two variants. (C) Example of LD between 300 genetic variants, calculated based on 450 samples on chromosome 22 within coordinates 10874444-15394184. Intensity of red colour indicates higher correlation (i.e. a higher R2 value, a parameter used to measure LD, with values closer to 1 indicating stronger correlation). The visually observed red triangles formed are typical in this kind of visualisation, representing the variants in specific regions (i.e. LD blocks) that are in high LD with one another. Panel (A) is inspired by educational lecture slides by Aaron Quinlan [17] and panels (B) and (C) by *Molecular Population Genetics* by Matthew W. Hahn [9]

1.2.2. Genetics of population structure

Understanding the genetic structure of human populations is essential for advancing knowledge in medical, forensic and anthropological sciences. A study by Novembre et al. [18] investigated the genetic structure of human populations using a sample of 3,000 European individuals. Researchers genotyped over half a million genetic variants to understand genetic variation in the European population. Despite low genetic differentiation among Europeans, a strong correlation was found between genetic and geographic distances. Consequently, a representation of Europe's geographic map has been created from the genetic data using Principal Component Analysis (PCA). The study highlights the need to account for genetic structure when researching disease phenotypes in order to avoid false associations. Furthermore, the findings show that an individual's geographic origin can be accurately inferred using their DNA, often to within a few hundred kilometres.

Access to large public human DNA databases, such as the 1000 Genomes Project [11], enables the use of PCA to determine the origins of samples with unknown provenance. This approach offers two key advantages: firstly, it provides an extra layer of quality control by comparing the reported and inferred population origins for any discrepancies; secondly, it helps account for genetic population structure in association analyses, preventing the identification of signals driven by population differences rather than the trait of interest.

1.3. Introduction to gene structure

A gene is a segment of DNA that contains the instructions for building a functional product, such as a protein or an RNA molecule. Genes are the fundamental units of heredity; they determine many of an organism's traits and characteristics, including physical traits like eye colour [19] or height [20] as well as biochemical traits like enzyme activity [21] or hormone production [22].

Genes are located on chromosomes, which are long, coiled strands of DNA found in the nucleus of eukaryotic cells (Figure 1). Each gene is made up of a specific sequence of nucleotides, which are the building blocks of DNA. The order of nucleotides in a gene determines the order of the sequence of nucleotides in an RNA molecule, which is later translated into amino acids in a protein. A combination of three nucleotides which code for an amino acid is called a codon, and each codon directs the cell to initiate the production of a protein chain (start codon), to append a distinct amino acid to the expanding protein chain, or to terminate the production of the protein chain (stop codon). There are 64 possible codons, which can code for 20 unique amino acids, whose combinations produce tens of thousands distinct proteins [23].

A single gene can be transcribed into multiple RNA molecules, which are known as transcripts, and these transcripts often produce the same protein. Almost every transcript consists of exons and introns, except a few single-exon genes [24]. Exons are the portions of the transcript that will remain in the

mature RNA (mRNA), whereas introns are removed by a mechanism called splicing during or after the transcription process is completed. Alternative splicing is a process that allows for the splicing together of various combinations of exons in a pre-mRNA molecule, which can produce multiple mRNA molecules from a single gene (Figure 3). Alternative splicing is not a completely stochastic process, but rather is governed by regulatory proteins and frequently depends on genetic variations within or in close proximity to the transcribed gene [25]. This mechanism introduces significant variability into genetic processes, making them even more complex and challenging to comprehend than before. [26–28]. While it is commonly misconceived that exons exclusively represent protein-coding regions of the transcript, in reality, only a fraction of exons (less than 30% on average in humans) are translated into proteins. The remaining exons reside in untranslated regions (UTRs) or non-coding RNA regions [29].

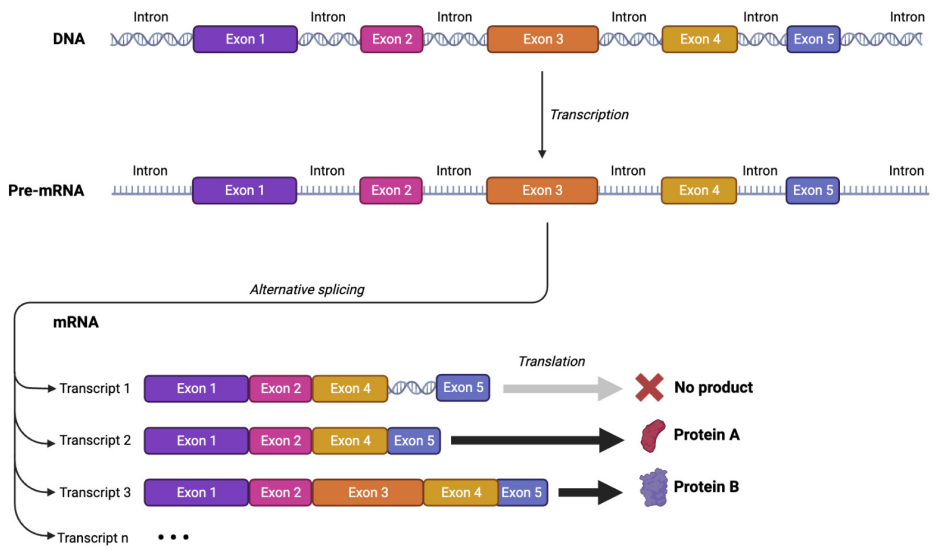

Figure 3: The function of alternative splicing in eukaryotic cell RNA processing: the gene on the DNA contains five exons. Once transcribed, the single-stranded pre-mRNA includes all exons and introns. Alternative splicing creates distinct mRNA versions: in the first transcript, exon 3 is omitted and the intron between exons 4 and 5 is retained, which resulted in the premature ending of the translation process, without protein production. In the second transcript, exon 3 is also excluded, while the third transcript undergoes constitutive (i.e. expected, non-alternative) splicing, where all exons are present and all introns are removed. These do not comprise a complete set of possible alternative splicing events and can be widened to include events such as alternative 5` and 3` splice sites, mutually exclusive exons and others. Following splicing, the mRNA molecule exits the nucleus of the cell, and if its sequence is suitable for translation, a protein is synthesised. The figure is created using the BioRender application (BioRender.com, 2023) by adapting available templates (i.e. RNA Processing in Eukaryotes, https://app.biorender.com/biorender-templates)

If you don't know where you want to go, then it doesn't matter which path you take

Lewis Carroll (Alice in Wonderland)

2. WHY DO WE CARE ABOUT MOLECULAR QTLS?

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

GWAS is a statistical method used to identify the genetic variants associated with a particular trait. The technique involves comparing the genomes of individuals with and without the trait of interest to identify the genetic differences that are more common in individuals with the trait. GWAS typically involves analysing millions of genetic variants across the genome, which can be time-consuming and computationally intensive. Traits of interest in GWAS can vary widely from phenotypes, which can be identified at the macro level (i.e. non-molecular phenotypes), such as height [20], body mass index [30,31], susceptibility to a certain disease [14], or even human intelligence [32,33], and micro level, such as metabolites [34,35]. Manhattan plots are commonly used to visually represent GWAS results (Figure 4).

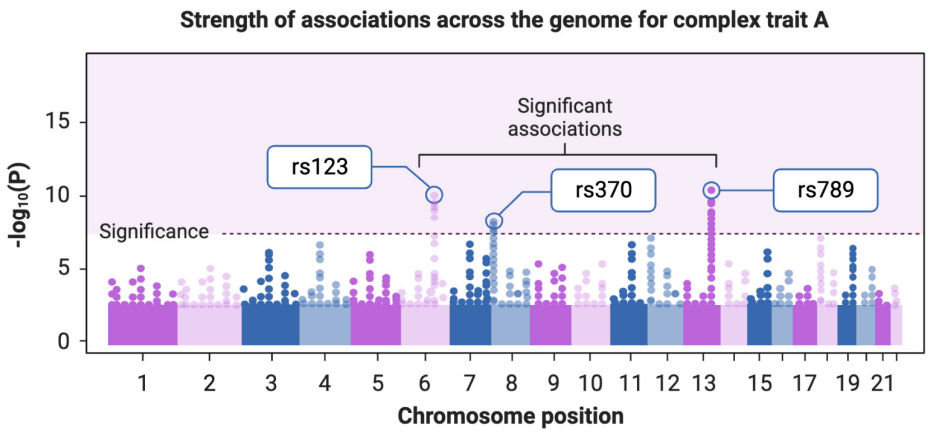

Figure 4: Visual representations of GWAS. Manhattan plots display the p-values for a complete GWAS (complex trait A in this example) across the genome. Each dot on the graph represents the p-value of the association between the trait and tested genomic variant at particular locus along the genome (x-axis). The value on the y-axis represents the $-\log_{10}$ of the p-value (equivalent to the number of zeros after the decimal point plus one). The figure is created using the BioRender application (BioRender.com, 2023) by adapting available templates (i.e. Manhattan Plot for Genome-Wide Association Studies (GWAS), https://app.biorender.com/biorender-templates)

Molecular Quantitative Trait Loci (molQTLs) are important in genetics because they help us to understand how genotypes relate to molecular phenotypes, such as expression of certain genes. By finding the genomic regions that affect gene expression or other molecular phenotypes (see Chapter 3.1.2.), molQTLs give us useful information about the mechanisms behind various diseases and complex traits when integrated with GWAS summary statistics. The expression of a molecular trait across the three possible genotypes can be

compared using a box-plot, providing a visualisation for a single molecular QTL (will be explained in more detail in chapter 3, Figure 6C).

Both GWAS and Quantitative Trait Loci (QTL) studies aim to identify the genetic variants associated with traits of interest. It is worth noting that the statistical models used in both GWAS and QTL analysis are identical (e.g. linear regression or linear mixed models). The key distinction between them lies in the fact that, in molQTL studies, the trait of interest can be located within the genome, allowing for the classification of associations as either *-cis* (from latin "on this side") or *-trans* (from latin "on the other side"), depending on whether the tested variant and molecular trait are located on the physically close or distant genomic regions, respectively [36]. As a consequence, this distinction leads to varying methods for data preparation and processing. In this chapter, we will discuss QTLs and their connections to GWAS, highlighting the ways in which these genetic approaches can complement each other.

2.1. GWAS – small effects waiting to be interpreted

GWAS finds associations between genetic variants and the trait of interest. However, GWAS does not elucidate the underlying molecular mechanisms responsible for the associations observed (Figure 5A). Without understanding the molecular mechanisms, it is hard to intervene in the process of expression of a trait of interest. For example, a polygenic risk score (a weighted sum of genetic variants usually derived from GWAS studies) for heart attack (i.e. myocardial infarction) [37] allows us to calculate the relative risk of having one; without understanding the molecular mechanisms behind the associated genetic variants, however, it becomes difficult to intervene or prevent the disease. To gain insight into the molecular mechanisms underlying a particular disease, one potential approach is to conduct QTL analysis to identify the genetic variants that affect the regulation of specific genes and determine whether these variants are also associated with an increased risk of developing the disease (i.e. GWAS) (Figure 5B). This approach can shed light on the complex interplay between genetics and disease and may ultimately lead to the development of new interventions to prevent or treat the disease [38].

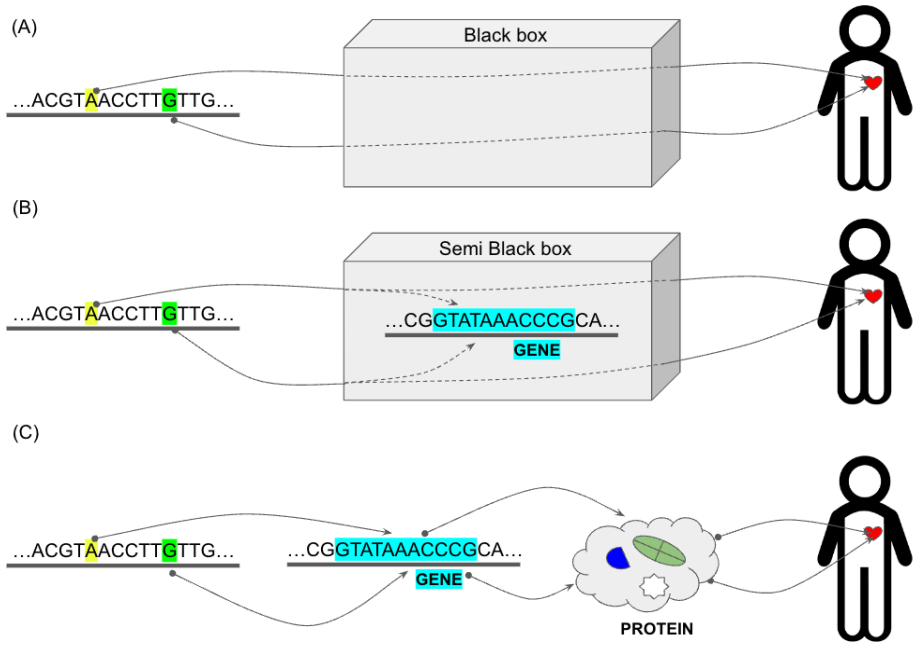

Figure 5: Using molQTL to gain insight into the possible molecular mechanisms causing the complex trait. (A) Representation of two genetic variants associated with a complex trait (e.g. heart disease). (B) Integration of QTL data into GWAS summary statistics to understand whether the genetic variants associated with the complex trait are also associated with the expression of certain genes. (C) Representation of a genetic variant impacting gene expression, which, in turn, affects the production rate of a related protein, ultimately influencing the complex human trait.

GWAS studies typically focus on common genetic variants that are present in a significant proportion of the population. More than 90% of GWAS hits are located in non-coding regions of the genome [39–42], such as regulatory regions that control the expression of nearby genes or regions that affect the three-dimensional structure of the genome [43]. Non-coding regions refer to regions of the genome that do not encode for proteins. Protein-coding regions (genes) make up only a small fraction (less than 2%) [44] of the entire genome [45]. When considering this property of GWAS in conjunction with the central dogma of molecular biology, it follows that a genetic variant can impact gene expression, which in turn affects the production rate of a related protein, ultimately influencing a human trait such as a disease (Figure 5C).

2.2. QTL – a very special form of GWAS

QTL analysis identifies the genetic variants associated with the variation of molecular quantitative traits and involves multiple steps such as data collection, phenotype quantification, normalisation and association testing. In QTL analysis we study associations between genetic variants and measurable molecular traits such as gene expression levels, which are typically quantified by using a technique called RNA-seq (will be described in more detail in subchapter 3.1). Each detected association between a genetic variant and a molecular trait is termed a molQTL, often simply referred to as a QTL. Some of the commonly used QTLs have specific names: eQTL refers to a gene expression QTL, while sQTL denotes a splicing QTL. QTL analysis is similar to GWAS in a broader context, as both of these approaches find associations between genetic variants and some kind of measurable property (e.g. phenotype) by applying linear regression. Technically, we can take expression level of gene A and calculate associations with variants in the whole human genome, put the p-value significance threshold at $5x10^{-8}$ (consensus genome-wide significance threshold, based on the assumption of having one million independent common variants) and call it a GWAS for gene A [46,47]. If we perform this process for all genes, we will ultimately be performing genome-wide QTL analysis. However, there are certain challenges and nuances to actually make it work. Firstly, the number of molecular phenotypes is often in the range of thousands (e.g. >20,000 for gene expression), and the number of common genetic variants in the range of millions (~10 million common variants in the European population, of which ~1 million are approximately independent (not in LD) from one another), which makes conducting comprehensive genetic analyses a computationally demanding task [48]. Secondly, it is known that genetic variants usually affect the genes that are situated nearby and located on the same chromosome (i.e. *-cis*) [49,50]. Hence, it is reasonable to perform this association testing in closer proximity to the phenotype of interest instead of testing against all genetic variants in the genome [51]. This approach not only simplifies the computational complexity but also allows for more accurate adjustment of the significance threshold, by using permutation approaches. Therefore, QTL analysis can be performed in *-cis* and *-trans*, representing associations in physically close or distant genomic regions, respectively (Figure 6).

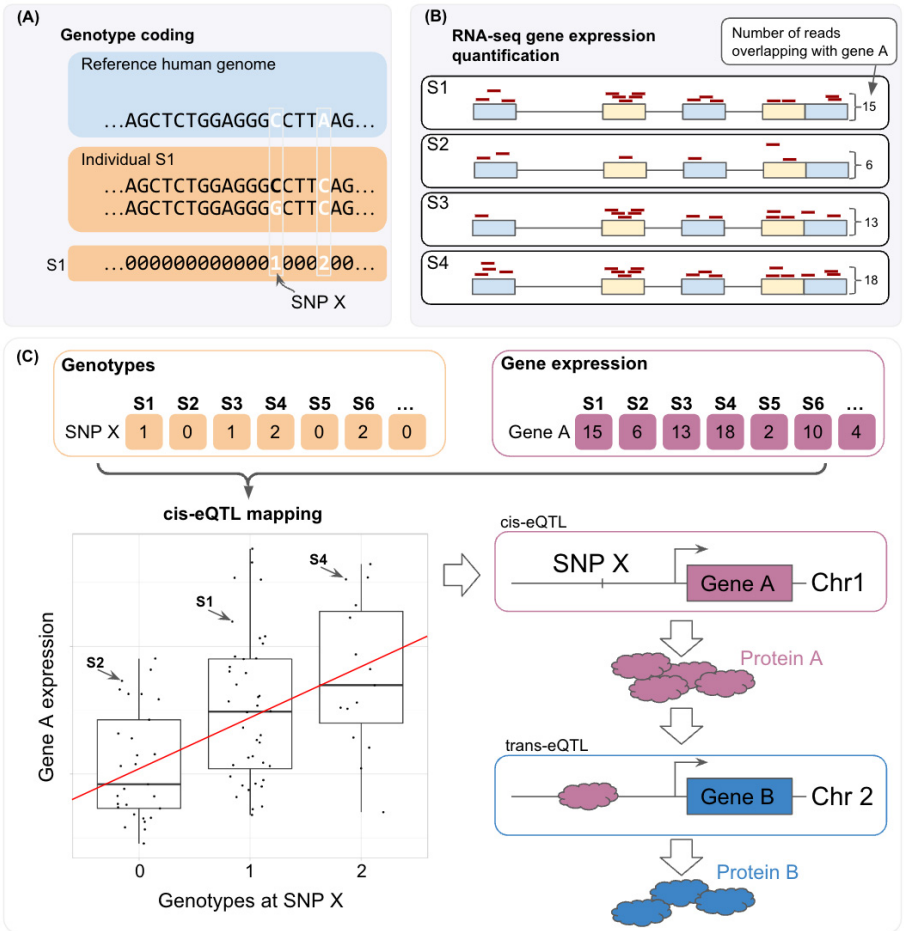

Figure 6: QTL mapping and visualisation example of one *cis*-eQTL (gene expression QTL) signal and a possible scenario of the trans effects of the same genetic variant. (A) Coding of genotypes in three possible combinations (i.e. 0, 1 or 2). This number indicates the total number of alleles of a genetic variant in an individual's genome that is different from the allele in the reference human genome. (B) Representation of gene expression quantification. For instance, sample 1 (S1) has 15 reads overlapping with gene A. In reality, these raw counts are normalised after quantification (see chapter 3.2); however, here, raw counts are used to simplify process representation. (C) The genotype data from (A) and gene expression data from (B) are used to map *cis*-eQTLs. *cis*-eQTL is visualised as a boxplot where there are three possible values of genotype on the x-axis and the corresponding gene expression value for each sample in QTL study is on the y-axis. In the current example, SNP X has an effect on the expression of gene A, and gene A codes for protein A. Both SNP X and gene A are located on chromosome 1 (Chr1). One good example of how *trans*-eQTLs can occur is when Protein A is a transcription factor (TF) protein and binds to the promoter region of gene B in another chromosome (e.g. Chr2), which in turn increases the expression of gene B and therefore also increases the production of protein B [52]. Visualisations are inspired by Figure 13 in [53] and Figure 1 in [52].

2.3. Using molQTLs to interpret GWAS hits

Studies have demonstrated that a significant portion of GWAS variants coincide with expression QTLs [43,54], suggesting that many disease-linked variants function by regulating molecular traits [55–57]. When a genetic variant is linked to both the increased risk of a certain disease and the expression level of a particular gene, it is possible that the variant influences the disease through that gene. By combining GWAS summary data and eQTL data, it is possible to identify target genes of disease-risk variants that cannot be detected using the GWAS method alone [58]. Additionally, eQTL analysis enables the functional characterisation of trait-associated variants through identification and prioritisation of the target genes in specific biological contexts, such as cell type and tissue. In Publication II, for instance, a colocalisation analysis between Vitamin D GWAS and eQTLs demonstrated that the GWAS variant alters Histidine Ammonia-Lyase (*HAL*) gene expression exclusively in the skin, rather than in any of the other 108 biological contexts examined.

2.3.1. Colocalisation

Colocalisation techniques are frequently used to examine whether the genetic variants associated with two distinct traits are consistent with a shared causal variant. These techniques involve comparing genetic associations from various datasets, such as GWAS and QTL studies, to determine the overlapping regions of genetic association. This approach can identify whether a single genetic variant is likely to be **causal** both for the trait and at the gene expression level, thus revealing the causal genes and regulatory mechanisms involved in complex diseases and traits. Previously, colocalisation methods have proven effective in detecting shared causal variants between molecular and disease traits within specific genomic regions [59–61]. COLOC is a popular colocalisation method that relies on summary statistics to estimate the odds of colocalisation by comparing five possible hypotheses (H0-H4) and calculates posterior probabilities (PP) for each hypothesis. Typically, the most interesting hypothesis in downstream analysis is H4, which states that associations with tested traits share one causal variant. Hence, if PP of H4 (PP4) is large, the data indicates that a single genetic variant causally affects both traits [62]. Nevertheless, a significant limitation of this method is its assumption that the region contains only one underlying causal variant per trait. Several novel ideas and tools have been proposed to overcome this limitation [61,63–65]. Chris Wallace's latest work proposes a more accurate colocalisation analysis method that allows for multiple causal variants. In this work, they adapted COLOC to utilise the Sum of Single Effects (SuSiE) framework [66], enabling multiple labelled comparisons in a genomic region and achieving higher accuracy compared to previous approaches, especially when more than one causal variant exists (Figure 7) [67].

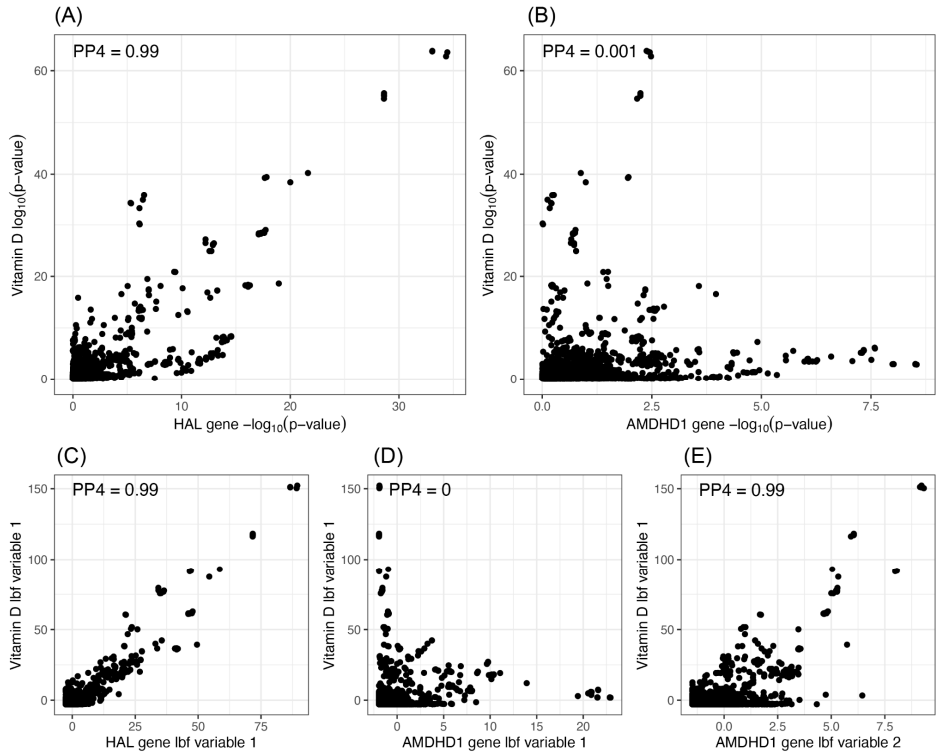

Figure 7: Comparison of COLOC V3 [62] and COLOC V5 [67] at the *HAL* locus associated with plasma vitamin D level in the UK Biobank. Log Bayes factor (lbf) variables represent the credible sets of the strongest associations. For example, lbf variable 1 represents the credible set which contains the strongest QTL, and lbf variable 2 represents the second strongest, etc. (A) Scatter plot of $-\log_{10}$ p-values of associated variants from Vitamin D GWAS and *HAL* eQTL in skin [68]. A high colocalisation posterior probability between the Vitamin D GWAS and *HAL* gene QTLs (PP4 = 0.99) suggests a shared causal variant is responsible for both traits. (B) Scatter plot of $-\log_{10}$ p-values of associated variants from Vitamin D GWAS and *AMDHD1* eQTL in skin. Colocalisation signal between Vitamin D GWAS and *AMDHD1* gene QTLs is weak (PP4 = 0.001), suggesting that there are two independent causal variants responsible for the two traits. (C) Scatter plot of first lbf variables of Vitamin D and *HAL* eQTL credible sets. Colocalisation signal between Vitamin D GWAS and *HAL* gene eQTL is high (PP4 = 0.99). (D) Scatter plot of first lbf variables of Vitamin D and *AMDHD1* gene credible sets. Colocalisation posterior probability between Vitamin D GWAS and *AMDHD1* gene eQTLs is low (PP4 = 0), as can also be seen in (B) the most strongly associated credible sets are different. (E) Scatter plot of the second fine mapped eQTL signal for *AMDHD1* and first lbf variable of Vitamin D. Colocalisation signal is strong (PP4 = 0.99). The second strongest association for *AMDHD1* colocalises with Vitamin D GWAS, but the strongest association does not. Hence, if we use COLOC V3, which assumes that the region contains only one underlying causal variant, we will miss the secondary colocalisation signal.

2.3.2. Fine-mapping

Fine-mapping is a statistical technique commonly used in genetics to identify the causal variants underlying a genetic association signal [69]. The technique is particularly useful in GWAS and eQTL analyses, where the high correlation between neighbouring SNPs can hinder the identification of potential causal variants. Fine-mapping involves partitioning the genomic region around significantly associated SNPs to independent regions and using various computational methods to infer which variants are most likely to be responsible for the association observed. One commonly used approach is based on LD patterns with the lead SNP, while Bayesian methods for fine mapping have become increasingly popular in recent years [70–74]. The result of fine-mapping is usually a minimal set of SNPs, known as a credible set, which capture the likely causal variant(s) with a certain probability (e.g. 95%).

These Bayesian methods compute the posterior inclusion probabilities (PIP) for each SNP as causal in a model and order the variants in decreasing order of PIP values. The minimal set of SNPs in the given region that captures the probable causal variant(s), known as a credible set, is then determined based on a pre-specified coverage probability threshold, usually 95%. The ranked PIP values are summed until the cumulative probability exceeds the given threshold; the corresponding top variants are considered to form a credible set. For example, the Sum of Single Effects Model (SuSiE) is one such fine mapping model [66,75].

2.4. Challenges in integration of GWAS and molQTLs

Despite the availability of tools for integrating molQTLs with GWAS to identify the molecular mechanisms underlying complex traits [62,63], the integration process remains nuanced. Firstly, QTLs are context-specific, resulting in varying levels of gene expression among different cell types, tissues and external stimuli, despite originating from the same DNA [76–79]. In contrast, GWAS is not context-specific, requiring the researcher to choose the appropriate QTL context for integration with GWAS. Secondly, the genetic variant's effect on a complex trait may not be attributed directly to gene expression or other molQTLs, but by different scenarios such as horizontal pleiotropy, linkage or reverse causality, making it challenging to identify or interpret the molecular mechanism responsible. These scenarios will be described in the next subsections. Finally, it is difficult to identify the precise molecular mechanism responsible for a complex trait due to the large number of potential molecular traits that could be involved. Despite these challenges, integration of GWAS and molQTLs can provide valuable insights into the genetic basis of complex traits.

2.4.1. Missing contexts (cell types, tissues, conditions)

All cells in the human body contain the same DNA code; however, despite having the same DNA code, cells differentiate during human development into different cell types, such as muscle cells, nerve cells, and blood cells, each with a specific function and characteristic gene expression pattern [80–83].

Lately, due to advancements in single-cell RNA-seq techniques, the precise count of human cell types has emerged as a subject of debate, as the discovery of "new" cell types has occurred [84–86]. Regardless, it is established that the human body comprises a variety of cells and tissues, with each type fulfilling a unique purpose [87,88], primarily controlled by the expression of various genes [89–91].

Gene regulation is the process by which cells control the expression of genes, which involves the activation or repression of specific genes at different stages of development or in response to environmental stimuli. Gene regulation is essential for the proper differentiation and function of different cell types and involves the interaction of multiple proteins, transcription factors and regulatory elements that can vary between cell types and tissues [92]. Gene regulation is a complex process; disruptions in this process can lead to various diseases, such as cancer and genetic disorders, underscoring the importance of understanding gene regulation in different cell types and tissues [93–95].

The selection of context for QTLs is crucial when integrating GWAS and molQTLs in order to identify the molecular mechanisms underlying complex traits. In order to achieve convincing colocalised signals, it is important to choose the most appropriate context, which can include cell type, tissue and other relevant conditions [79]. While some complex traits may have predictable contexts that are most suitable for integration; in many cases, the optimal context may not be obvious or the required QTL data may not be available. As an example, suppose we have GWAS summary statistics and we wish to detect colocalising QTL signals in order to identify target genes explaining the molecular mechanisms of the complex trait. Despite the availability of QTL summary statistics resources such as those from the Genotype-Tissue Expression project (GTEx) [51], eQTL Catalogue [96] and eQTLGen [48], it is still possible to overlook the QTLs that could explain the molecular mechanism underlying the GWAS trait if the corresponding QTL study in the relevant context is not available. In such cases, even though colocalisation tests can be performed with all available contexts, the lack of QTL data for the desired (potentially unknown) context can result in missed opportunities in identifying important molecular mechanisms. Furthermore, colocalisation can lead to false positive or deceptive outcomes when examined in an incorrect context. In such contexts, the causal molQTLs may be absent, resulting in the detected colocalisations (albeit individually uncommon) being prone to false positives, such as when two independent causal variants are in linkage disequilibrium. A study by Mostazavi et al. [97] developed a model to systematically compare the GWAS and eQTL variants and suggests that even in QTL studies with large sample

sizes, the identified eQTLs tend to favour enhancers functioning in less restricted cell types or environmental settings that hold less significance for interpreting GWAS variants.

2.4.2. Exploring effects of genetic variants on complex traits

Colocalisation analysis is usually performed under the assumption (or hope) that a genetic variant affects a complex trait by altering gene expression (Figure 8A). However, it is also possible for a genetic variant to be associated with both a complex trait and gene expression through alternative mechanisms, including horizontal pleiotropy, linkage and reverse causality [98].

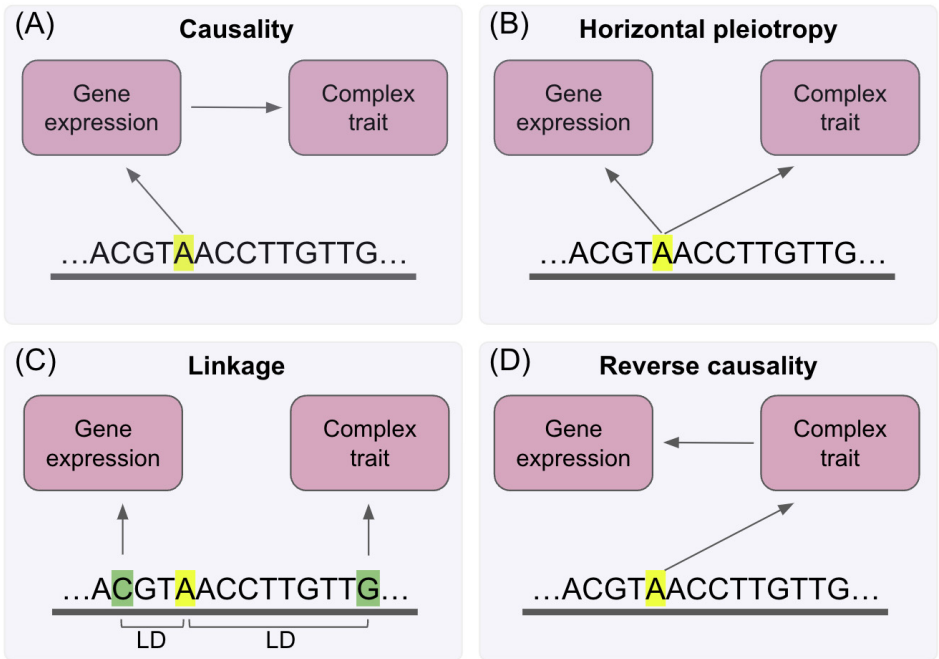

Figure 8: Different pathways of genetic variants affecting the complex trait. (A) Causality: when the genetic variant affects the expression of a gene, and the expression of a gene affects a complex trait. (B) Horizontal pleiotropy: genetic variant affects the gene expression and the complex trait independently. (C) Linkage: two genetic variants which are in high LD with the tested variant affect the gene expression and the complex trait independently. (D) Reverse causality: genetic variant affects the complex trait, and the complex trait affects the expression of the gene. This figure is inspired by Figure 7 of Kaido Lepik's doctoral thesis [98].

Horizontal pleiotropy refers to a situation where a genetic variant influences multiple traits that are not part of the same biological pathway. In other words, a single genetic variant affects multiple phenotypic traits independently of each other (Figure 8B). Indeed, this can result in the manifestation of a colocalisation signal in both GWAS and QTL studies; however, it does not necessarily imply

that the genetic variant influences the trait through the regulation of gene expression. Horizontal pleiotropy can be a confounding factor in colocalisation analysis and needs to be accounted for. Other statistical methods, such as Mendelian randomisation, can be used to try and account for horizontal pleiotropy [99,100] and assess the causal relationship between the genetic variants, gene expression and the complex traits [101–103].

Linkage disequilibrium can have implications for GWAS and molQTL analysis. For example, if an associated genetic variant in a GWAS is in LD with two other variants that independently affect the molecular trait and the complex trait, it may appear as though the original variant is affecting both the complex trait and the molecular trait (Figure 8C).

Sometimes, it is possible that the link between a genetic variant and gene expression is due to the complex trait rather than the gene expression causing the trait. This can happen because the expression of certain genes in a relevant cell type might change when a complex disease occurs, leading to the assumption that the gene expression caused the disease when in fact it was a result thereof [104,105]. Although it is more relevant in *trans* QTL studies, and much rarer in *cis*-QTL studies [48,106], it is important to keep in mind the possibility of this scenario (Figure 8D, Figure 6).

2.4.3. Challenges in defining the molecular mechanism (splicing vs. expression)

The most commonly used QTLs are gene expression QTLs (eQTL). In the eQTL Catalogue, we have generated and shared QTLs for another four quantitative traits (namely, exon expression, splice-junction usage, transcript usage and transcriptional event usage). However, QTL analysis can be performed for a wide range of molecular quantitative traits, including but not limited to histone modifications [107,108], chromatin accessibility [109,110], alternative polyadenylation [111,112], and DNA methylation [113]. Therefore, even if we identify the appropriate QTL context to colocalise with GWAS, it is still possible that the chosen quantitative trait may not be sufficient for explaining the underlying molecular mechanism. It is helpful to note that chromatin, histone and DNA methylation QTLs do not directly reveal the target gene. Identifying target genes (and thus potential mechanisms) from these colocalisations still requires the use of eQTLs or alternative methods. Additionally, it is known that different types of molQTLs are usually shared [96]. For example, a strong eQTL can also be manifested as a weak splicing QTL, or the other way around. This makes it even more difficult to pinpoint the precise sequence of events leading to the occurrence of a complex trait.

I'm a greater believer in luck, and I find the harder
I work the more I have of it

Thomas Jefferson

3. QTL ANALYSIS

QTL analysis (also referred as QTL mapping) involves two main inputs: genotype data and molecular trait data. While standard approaches have been developed for genotype data processing and imputation [114,115], a considerable amount of effort is needed to quantify molecular phenotype data. In the following subchapters, we will discuss the steps involved in quantification, quality control (QC) and normalisation of data as well as the methods used in QTL mapping and the interpretation of QTL signals (Figure 9).

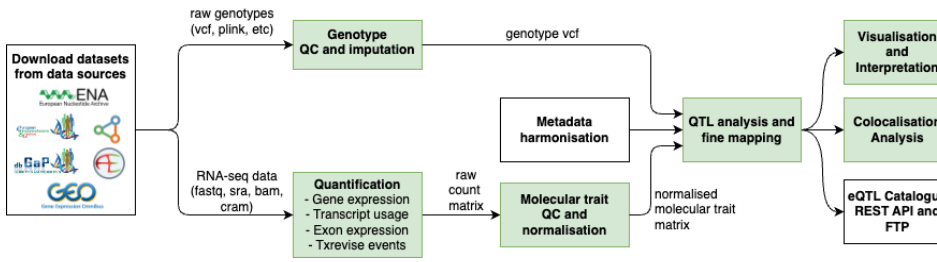

Figure 9: High-level representation of eQTL Catalogue workflow. First, the necessary data requests are made and permissions are acquired. Then, the genotype data are quality controlled and (if needed) imputed using the *genimpute* pipeline [115]. RNA-seq data are quantified with the *rnaseq* pipeline [116] and quality controlled and normalised with the *qcnorm* pipeline [117]. Then, the outputs of *genimpute* and *qcnorm* pipelines and harmonised metadata are gathered together and fed into the *qtlmap* pipeline [118] to perform association testing. The outputs of the *qtlmap* pipeline (i.e. association testing summary statistics) are made available via the eQTL Catalogue FTP server and an application programming interface (API). There are two optional downstream analysis pipelines, *coverage_plot* [119] to visualise and interpret molQTLs and *colocalisation* to perform a colocalisation [120] analysis with other association studies like GWAS. The steps in the workflow featuring a green background represent those for which a specific pipeline has been established.

3.1. Quantification of molecular traits

Comprehending the meaning of “molecular trait” can often be challenging as it encompasses any measurable entity at the molecular level. Nevertheless, in this thesis, the term “molecular trait” generally refers to a quantified measure of messenger RNA (mRNA) and its isoforms. Due to the intricate nature of RNA biology, there exist various means of assessing the abundance of mRNA products. Regardless, understanding the origin and form of a signal is crucial when quantifying any type of signal. Consequently, to measure molecular traits accurately, it is necessary to become acquainted with the RNA sequencing (RNA-seq) process that produces the data and the data itself.

3.1.1. RNA sequencing (RNA-seq)

RNA-seq is a powerful tool for studying gene expression and the transcriptional landscape of biological systems. RNA-seq involves several key steps, beginning with mRNA extraction from a biological sample. The extracted mRNA is then processed to create a sequencing library, which involves converting the mRNA to complementary DNA (cDNA) using reverse transcription, fragmenting the cDNA into small fragments and adding adapters to the fragments [121]. Following library preparation, high-throughput sequencing technologies such as Illumina sequencing are employed to generate millions of short reads. Typically, these reads range from 50 to 500 base pairs in length and are aligned to a reference genome or transcriptome to determine the specific genomic position in which each read originated [36]. RNA-seq output is typically provided in either the FASTA or FASTQ file format. Fasta files contain a sequence header line, which begins with a ">" symbol, followed by the sequence data on the next line(s). Each header line represents a unique sequence, and multiple sequences can be stored in a single fasta file. FASTQ files also contain additional information about the quality of each base in the sequence, which is not present in FASTA files (Figure 10). In FASTQ files, each sequence is represented by four lines: the first line contains the sequence header, the second line contains the nucleotide sequence data, the third line contains a "+" symbol, and the fourth line contains the quality scores for each base in the sequence (Figure 10). To reduce disk space usage, FASTQ files are often compressed using gzip, resulting in a file extension of .fastq.gz. For RNA-seq libraries that have been sequenced using a paired-end protocol [122], FASTQ files are generated in pairs with file extensions that match the format _R1.fastq.gz and _R2.fastq.gz.

```
@HWI-ST301L:302:C1BW3ACXX:6:1101:1739:2182 1:N:0:TAGCTT
AAAGGTCAGGTCAATGTCCTAATGTCAGCAACTCTGTGCAGTAGAATCACAGGGTGGCTTTGTAAGGAGCCACGTGCTTAGGACCCACTCCAGACCTCTG
+
@CCFFFDFFDHHGJJIJAHHIIHJCGGGIHGGGGGJJ*CCGGCH><DGIII?B@BFGI>FGEIJBH<FGDDHFEEBB=C>>>;;?;?BA=A5::?BCCD5
```

Figure 10: Example of an RNA-seq read from FASTQ file.

3.1.2. Defining molecular traits

The quantification of RNA-seq data involves several steps, including read alignment to the reference genome, transcriptome assembly, and the determination of gene expression levels. Each of these steps has multiple algorithmic options, with researchers selecting the most appropriate tool based on their specific experimental design and research question.

The ultimate outcome of quantifying molecular phenotypes is the list of examined phenotypes with the expression level metric given per each sample. An example of such an outcome is given in Table 1.

Table 1: An example of a quantification result in tabular numeric format. The identification of the molecular phenotype is given in the initial column, while all subsequent columns represent the corresponding quantity of the molecular trait in each sample.

phenotype_id	sample_1	sample_2	sample_3	sample_N
ENSG00000223972	7	0	222	4
ENSG00000227232	0	22	1	422

While there is no clear consensus on which RNA-seq quantification method is optimal, the continued development of new tools and algorithms has led to improvements in the accuracy and reproducibility of RNA-seq measurements. Moreover, the increasing availability of public RNA-seq datasets and tools has facilitated the benchmarking and validation of RNA-seq quantification methods [123–126]. Despite ongoing debates, the quantification of RNA-seq data remains a crucial method in molecular biology, enabling researchers to investigate gene expression, alternative splicing, and other RNA-based phenomena in a variety of biological contexts.

In the eQTL Catalogue [96,127], as outlined in Publications I and II, five transcriptional phenotypes were quantified (Figure 11). The following subchapters will examine the tools currently available for quantifying relevant molecular traits. Keeping the analysis context in mind is crucial, since all the decisions made are aimed at preparing the data for utilisation in QTL mapping.

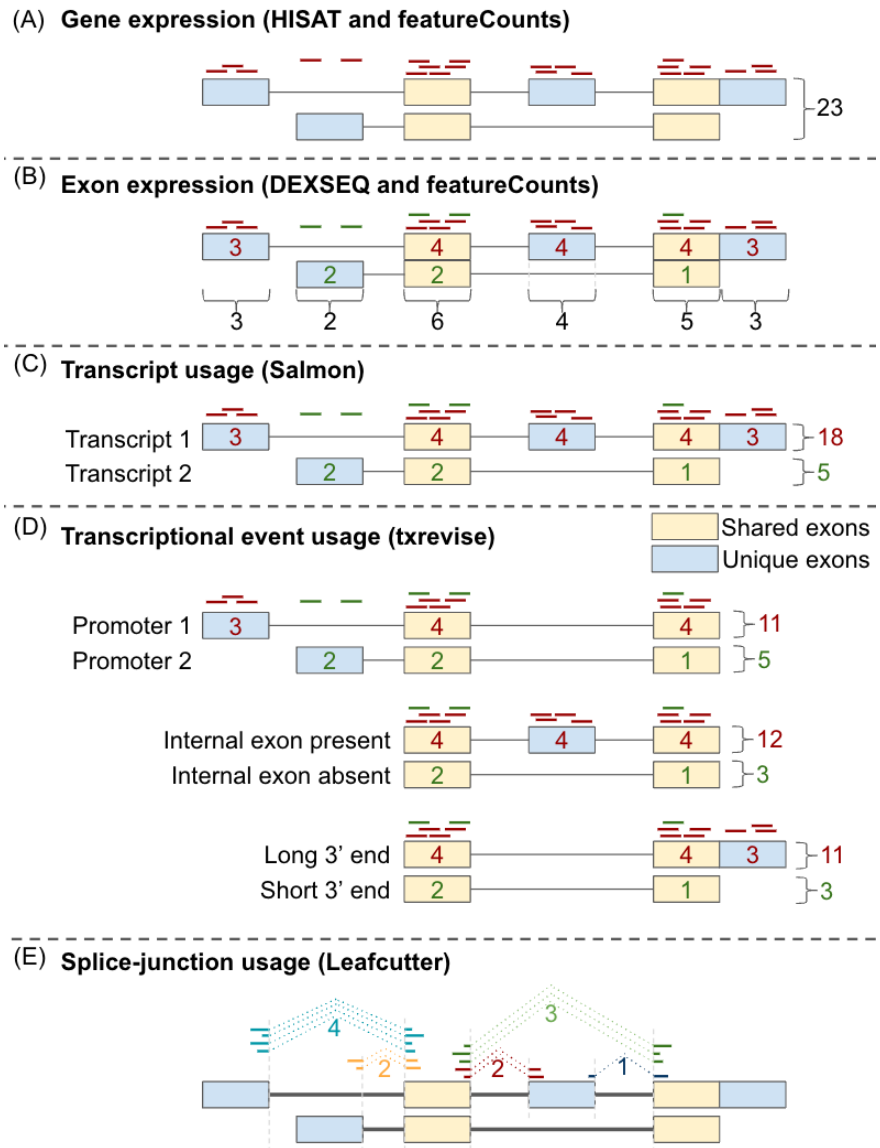

Figure 11: Overview of the five molecular trait quantification methods used by the eQTL Catalogue. (A) Gene expression was quantified by counting the total number of reads overlapping annotated exons of the gene. In this example, the gene has 23 overlapped reads. (B) Exon expression was estimated by counting the number of reads overlapping each exon. In example, the first exon has three and the second exon has two overlapping reads. (C) Transcript usage was estimated with Salmon. In the example, the first transcript has 18 and the second transcript has five assigned reads. For better understanding, the total number of assigned reads to transcripts and exons is 23, exactly the same as the number of reads overlapping the gene. (D) Txrevise was used to estimate the expression levels of three types of transcriptional events (promoter usage, splicing and 3′ end usage). (E) Splice-junction usage was quantified with Leafcutter by counting

the reads that overlap with splice-junctions of two exons. Rectangles represent exons and the horizontal black lines connecting the rectangles represent introns. One row, which consists of exons and introns, represents a transcript of a gene. The small horizontal non-black lines overlapping the exons represent the reads. The original figure is published in the Supplementary Figures of Publication II and has been adapted for inclusion in this thesis.

3.1.2.1. Gene expression

The expression of a gene is merely a representation of the quantity of mRNA molecules originating in that particular gene (Figure 11A). To measure gene expression, we initially align the short RNA-seq reads to the reference genome. The outcome of the read alignment process produces Binary Alignment Map (BAM) files that contain individual records indicating the genomic coordinates to which the alignment software assigns each read. These files can be further analysed through visualisation software, such as Integrative Genomics Viewer (IGV), which allows for exploration of the genomic regions to which the reads align (Figure 12). There are several tools for doing the RNA-seq read alignment [128], but the most popular ones are STAR [129] and HISAT2 [130]. To achieve fast alignment, STAR utilises suffix arrays. However, to function properly, it demands a considerable amount of random access memory (RAM) (~27 gigabytes). In contrast, HISAT2 utilises an indexing approach based on the Burrows-Wheeler transform and the Ferragina-Manzini index, which requires less memory. Benchmarking studies have shown that STAR and HISAT2 perform to a similar accuracy, and correlation of raw count distributions are quite high (>99%) [128].

Once the RNA-seq reads have been aligned, they are then assigned to specific genes to determine the degree of overlap between each read and its corresponding gene. Although it may seem like a simple task to count the reads that overlap with genes in a specific genomic position, it is actually quite challenging. First, there may be more than one gene located at a single position (i.e. multi-overlapping reads), including overlapping genes on the opposite strand of the DNA. Second, the reads could be mapped to multiple locations within the genome (i.e. multi-mapping reads), resulting in ambiguity regarding the gene they should be assigned to. In other words, the ambiguity around the origin of the read can make it difficult to accurately count the reads that overlap with specific genes. To address these challenges many software tools have been developed [123,131–133].

Our decision to use HISAT2 instead of STAR is mainly motivated by computational efficiency while maintaining mapping accuracy at a reasonable level [128]. When it comes to basic read counting, the majority of tools exhibit similar performance to featureCounts [128,134]. Another approach involves quantifying gene expression by aggregating transcript-level counts per million fragments (TPMs) from Salmon [133] or Kallisto [132]. While this method may offer potential accuracy advantages, it could diminish eQTL interpretability if the reads align to several genes.

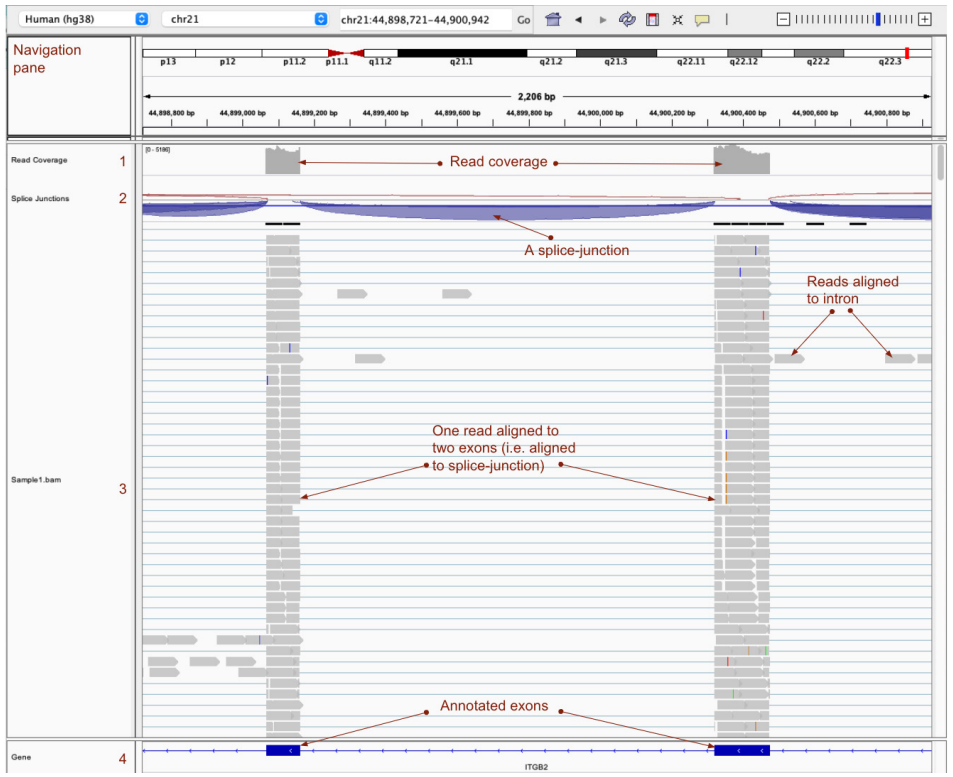

Figure 12: A screenshot from IGV software. A BAM file is loaded to the IGV and shown together with human genome annotation. At the top of the image is a navigation pane showing the currently visible region of the genome. In this example, we are seeing the region 44,898,721-44,900,942 of chromosome 21, where we can explore two exons of *ITGB2I* gene. In the column positioned on the left, the titles of four tracks are shown. Read Coverage shows the histogram of the reads aligned with the corresponding position. Splice Junctions showing the reads aligned with at least two exons. Sample1.bam shows the individual reads aligned with the corresponding positions. The last (bottom) track "Gene" shows the annotation in the selected region. When the reads are aligned with more than one exon (i.e. aligned with a splice-junction), a horizontal line connects them.

3.1.2.2. Exon expression

Not all the base pairs from gene start to end are used to produce proteins. Typically, a large portion of the gene consists of non-coding regions, such as introns, which will eventually be excluded via machinery called splicing [27,135]. The process of quantifying expressed exon counts is quite similar to gene expression quantification. First, the reads are aligned to the reference genome, and then the number of reads that overlap with the trait of interest (in this case, exons) is counted (Figure 11B). However, in order to account for each possible exon, it is necessary to flatten the gene model when the exon boundaries vary across different transcripts of the same gene. We utilised specialised software called

DEXSeq to create the exon annotation file from scratch [136]. Within this annotation, when there are two or more overlapping exons at a locus, they are separated at the intersection sites to create exonic parts (or as mentioned in the original DEXSeq article, counting bins). This approach allows for the counting of reads that overlap with a single "exon" in a specific locus, as the possibility of overlapping with multiple exons is eliminated (Figure 13). Sometimes, the size of the overlap is too small—only a couple of base pairs—which results in a very short exonic part. While these short exonic parts may lack biological significance, we find it acceptable to use them as a solution to the even more complex task of counting reads that overlap with multiple exons.

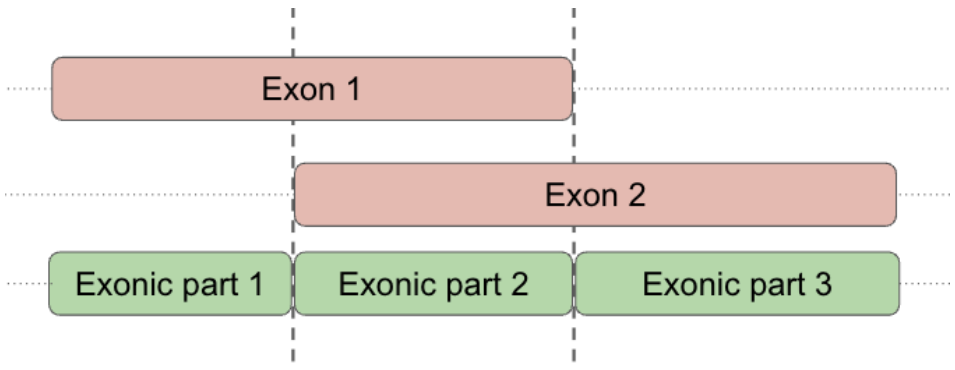

Figure 13: Generation of exonic part annotation. When there are two overlapping annotated exons, they are separated at the intersection sites to create exonic parts. In the figure, Exon 1 and Exon 2 are overlapping; hence, in the new exon annotations, three exonic parts are created from these two exons.

3.1.2.3. Transcript usage

As previously noted, genes can be expressed in various isoforms known as transcripts, which are produced through alternative splicing [28]. While multiple transcripts can arise from a single gene, they usually produce the same product, typically a protein. Quantifying transcript "expression" within the context of a single gene is slightly different from quantifying gene and exon expression (Figure 11C). Even though it is possible to measure the absolute expression of transcripts by counting overlapping reads, similar to the process for genes and exons, this may not be particularly insightful. This is because the total expression of a gene's transcripts will simply equal the expression of the gene itself. Instead, in the context of a gene being expressed, it is more informative to examine, which of its transcripts are used relatively more often than the others (i.e. usage). Thus, if we consider the gene expression level of a certain gene to be constant for a given sample, high usage of one transcript will lead to low usage of all other transcripts of that gene. Furthermore, examining this relative usage can help to gain new insights into different molecular mechanisms. Hence, in this context, the term "usage" is used for transcripts rather than "expression" to refer to the relative property of transcripts within a gene.

The advantage of transcript usage quantification is that it does not require precise base-to-base alignment of reads to the reference; rather, it is sufficient to identify possible locations of origin for the read. Multiple tools have been developed to benefit from this property, such as RSEM [137], Kallisto [132], Sailfish [138] and Salmon [133], and benchmarking experiments have compared these tools [126]. We decided to use Salmon for its speed and the ability to perform transcript quantification with various custom-made transcriptome annotations, as described in the Transcriptional Event Usage section below. Salmon is a quantification method for transcript abundance estimation in genomics research that first uses quasi-mapping [139] to map the reads to the transcriptome, then uses a dual-phase statistical inference procedure and sample-specific bias models to account for technical biases. It provides a probabilistic model of the sequencing experiment, which incorporates information like the terms contributing to the conditional probability of drawing a fragment of a given transcript. It achieves the same order-of-magnitude benefits in speed as Sailfish and Kallisto but with greater accuracy [133]. These features make Salmon an efficient and accurate tool for transcript abundance estimation, particularly in transcript usage QTL mapping analysis, where the assignment of reads to the wrong transcripts may have important implications.

3.1.2.4. Transcriptional event usage

It is essential to assume that all expressed transcripts are included in the transcriptome annotation for accurate quantification of transcription usage. If any expressed transcripts are missing from the reference transcriptome, it could result in the reads from those missing transcripts being wrongly assigned to other (non-missing) transcripts that are not expressed, leading to inaccurate quantification [140]. Unfortunately, most of the alternative promoter and 3′ end usage-related transcripts are missing from commonly used reference transcriptomes [141,142]. MISO was created for performing an 'event-level' analysis that splits reference transcripts into individual events and then estimates the expression of each event using conventional transcript quantification techniques [143]. Nonetheless, MISO addresses only a portion of promoter events, specifically alternative first exons, and the event annotations have not been revised in a considerable period of time. To address this, Alasoo et al. developed a complementary approach called “txrevise” to overcome limitations of MISO by stratifying reference transcript annotations into separate events, namely promoter, splicing and 3' end events (Figure 11D) [79]. It has later been expanded to include the experimentally identified Cap Analysis of Gene Expression (CAGE) [144] promoters from the FANTOM5 project [145] into its annotations [146]. Practically, using txrevise is a simple process. Instead of using the reference transcriptome for transcript usage quantification, custom-made transcriptional event annotations are utilised as input to build a Salmon index, after which Salmon is used to quantify transcriptional event usage.

3.1.2.5. Splicing junction usage

To investigate the associations between genetic variants and splicing events, it is necessary to first quantify these splicing events. Various tools have been developed for quantifying alternative splicing, including those that rely on transcript usage [132,138,147,148], exon inclusion levels [136] or transcriptional event usage (MISO, txrevise). However, all of these approaches are dependent on transcript models, which may be inaccurate or incomplete [149].

To address this issue, we incorporated Leafcutter [150] into the eQTL Catalogue. This tool is specifically designed to detect removed introns by counting the reads that align to exon-exon junctions, rather than relying on pre-defined splicing events or transcript models (Figure 11E). If the molecular traits of interest are specific isoforms of a gene (i.e. transcript) or exons of an isoform, they can be easily classified as belonging to a particular gene. However, it is not always clear how to assign a splice junction event to a specific gene. To address this issue, the authors of Leafcutter proposed a useful method for grouping these junction usage events into phenotype groups, known as clusters, by assigning reads with the same exon-exon junction to the same cluster. In order to assign clusters (and all the splice junctions within it) to gene(s), we check whether the cluster shares at least one overlapping splice junction with the beginning or end of an annotated exon. Phenotype groups can be utilised in the permutation testing process (i.e. multiple testing adjustment) during QTL mapping.

3.2. Quality control and normalisation

Quality control (QC) is a critical step in any data analysis project as it ensures the accuracy and reliability of the results. Without proper QC, the data can be misleading, resulting in incorrect conclusions and wasted resources. The methods sections of Publications I and II and my master's thesis [96,127,151] thoroughly explain all quality control related aspects. In summary, the QC procedures we utilised allowed us to recognise and eliminate problematic samples from the eQTL Catalogue. These procedures comprised several steps, including identifying outliers via principal component analysis (PCA) and multidimensional scaling (MDS) analysis; performing a Match BAM to VCF (MBV) analysis which checks for genotype concordance between RNA-seq BAM files and genotype VCFs and flags sample mixups and potential cross-contamination; conducting gene expression analysis that accounted for sex-specific differences; and eliminating specific gene categories (such as pseudogenes and short RNAs that are not reliably quantifiable with short-read RNA-seq) or genes with exceptionally low expression levels.

Normalisation is also a crucial step in QTL analysis as it reduces the effect of technical variations on the results, allowing for the more accurate and reliable identification of true biological variation. Failure to normalise the data properly can lead to false signals, especially false positives that are likely to be driven by outliers, which can compromise the interpretation of the results (Figure 14).

Additionally, normalising the data helps to ensure that the distribution of the quantified molecular trait residuals is close to normal, which is a requirement for linear regression used in QTL analysis. Inverse Normal Transformation (INT), which we apply in the normalisation step of all quantification methods, significantly reduces the impact of individual outlier samples that might be randomly correlated with a genetic variant.

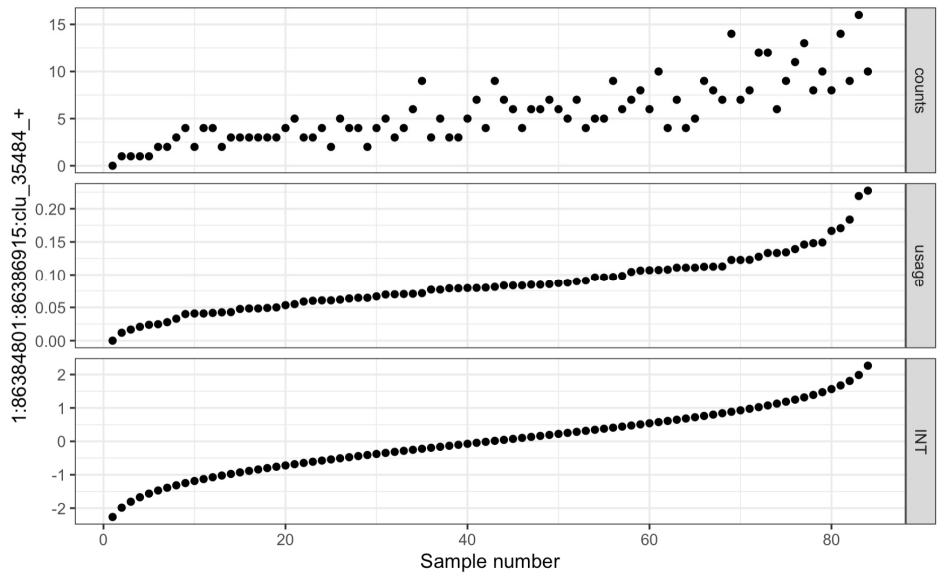

Figure 14: Before and after the normalisation process. The upper panel represents the raw counts of reads overlapping with the *1:86384801:86386915:clu_35484_+* splice-junction in macrophages that are stimulated with interferon-gamma [152]. The middle panel "usage" is extracted by dividing the raw counts by the total number of the reads assigned to the corresponding group (e.g. *clu_35484_+*) for each sample. Then it can be normalised with inverse normal transformation, as shown in the lowest panel.

3.3. Genotype data preparation

Preparing the genotype data is a challenging task that requires careful attention. The need for genotype imputation may vary depending on the technology used to generate the genotype data. In the eQTL Catalogue project, the genotype imputation-related tasks (i.e. pipeline development and execution) were performed by other members of the eQTL Catalogue team (See Chapter 5, Publication I). We did not perform imputation for studies that utilised Whole Genome Sequencing (WGS) technology but we did for studies that used array-based genotyping. Irrespective of the genotyping method employed, a rigorous quality control procedure is applied to all genotypes. After applying several filters, including removing variants using the Hardy-Weinberg equilibrium $P < 10^{-6}$, missingness > 0.05, and MAF < 0.01, we also removed samples that had more than 5% of their genotypes missing (Kerimov et al., 2021). After

imputation, we further excluded variants with an imputation quality score less than 0.4 from the imputed genotypes. These exclusions are implemented to decrease the likelihood of identifying spurious associations.

3.4. QTL mapping

In essence, QTL mapping involves identifying correlations between two datasets: the normalised and quality-controlled molecular trait matrix (quantified from RNA-seq data using a quantification pipeline) and the genotype data (Figure 6). The earlier subsections, 3.1 and 3.2, provide a comprehensive discussion on the molecular trait data and subsection 3.3 on the genotype data preparation.

3.4.1. Covariate adjustment

In our literature review, informed by the benchmark conducted by Zhou et al. [153], PCA surfaced as a more favourable method compared to other covariate adjustment techniques such as Surrogate Variable Analysis (SVA) [154], Probabilistic Estimation of Expression Residuals (PEER) [155,156], and Hidden Covariates with Prior (HCP) [157]. Notably, PEER, which employs a Bayesian probabilistic model, initialises with the PCA solution, and the factors derived are almost identical to PCs, as evidenced in GTEx eQTL and sQTL data [153]. On the other hand, SVA involves a complex algorithmic iteration between reweighting features of the molecular phenotype matrix and performing PCA on the resultant matrix. HCP also bears resemblance to PCA, as it minimises a loss function akin to the minimum-reconstruction-error loss function of PCA, albeit through a more complex coordinate descent optimisation. These methods can be seen as extensions of PCA, but the complexity they introduce proved to be a burden rather than a benefit in our case. Significantly, both PEER and SVA demand substantially more resources to execute, given the extensive number of datasets (500+) we are handling. Unlike GTEx [51], which optimised factor numbers for PEER with two quantification methods, our study encompassed three additional quantification methods—making factor optimisation computationally expensive. Furthermore, PEER and HCP lack straightforward methodologies for determining the optimum number of factors (K), often resorting to suboptimal strategies like maximising the number of discoveries to choose K, which would add computational burden on our pipeline. In light of these considerations, we opted for a simpler approach, employing a small fixed number of PCs to adjust for major batch effects without over-correction. This approach, while being cost-effective in terms of computation, also removed the need for extra quality control steps that more complex methods like PEER would require, particularly when used for ratio-based quantification methods such as transcript usage and splice junction quantification, for which they were not initially designed.

3.4.2. Adjusting for multiple testing

Our analysis necessitates a two-fold correction: firstly, for the number of variants in the -*cis* region of each gene, and secondly, for the overall number of genes under consideration. The fact that variants around a gene are not independent due to linkage disequilibrium (LD), rendering traditional methods like the Bonferroni correction excessively stringent and thereby unsuitable for this context. An alternative approach, eigenMT [158] accounts for LD among variants by estimating the approximate number of independent tests performed. An additional complexity is that for transcript-level quantification methods, we should also explicitly account for the number of individual events tested per gene, which is tricky to do with methods such as eigenMT. To navigate these intricacies, we have chosen to employ permutation testing [159] to adjust both for the number of variants for each gene as well as the number of events tested per gene. This empirical permutation approach elegantly sidesteps the assumptions tied to other techniques while still providing robust correction. Following this, we apply a Benjamini-Hochberg False Discovery Rate (FDR) of 1% to adjust for the multiple testing across all tested genes.

3.4.3. Difficulties in QTL mapping

While the theoretical concepts of QTL mapping are firmly established [36], the current difficulty in conducting this analysis on a large scale lies in the accessibility of the data and the technological framework required to handle diverse contextual QTL investigations uniformly. The primary aim of this thesis is to disseminate the best practices and knowledge gained from constructing scientific pipelines to overcome the technical challenges encountered in QTL analysis. Although the major drawbacks of developing scientific pipelines were discovered during the creation of the QTL analysis pipeline, the insights garnered can be extended to any form of scientific analysis that entails handling extensive amounts of data. This thesis will comprehensively cover the fundamental principles of developing scientific pipelines in Chapter 4.

3.5. Post-processing, sharing and interpretation of QTLs

Studying QTL data in isolation is not typically effective for uncovering molecular mechanisms related to complex traits. For a more comprehensive understanding, it is necessary to integrate these data with other information, such as signals from GWAS (Figure 5B). There are several ways to use QTLs for the interpretation of molecular mechanisms of a complex trait. A common approach involves fine-mapping variants to identify conditionally independent signals and performing colocalisation analysis with GWAS to identify potential candidate genes related to the complex trait. However, interpreting colocalisation results can still be challenging due to the presence of multiple correlated variants in the colocalised locus and potential horizontal pleiotropy. Moreover, the summary statistics alone do not provide a clear assessment of the magnitude

and direction of the genetic effect, the affected region within the gene, or the absolute expression of the affected transcript, especially for transcript-level and splicing associations. These ambiguities can often be resolved by examining RNA-seq read coverage and exon-level QTL visualisations together with the QTL of interest [127].

Upon completion of QTL analysis, it is important to publicly share the summary statistics for others to use, although the size of the data and the effort required to prepare it for sharing are often underestimated. In the initial release of the eQTL Catalogue [96] the shared data size exceeded 15 terabytes, and we provided three solutions for researchers to access the generated summary statistics: downloading the entire dataset using FTP, fetching specific regions of the file using Tabix based on chromosome and positions of interest [160], or using an API to request specific types of data, such as associations for a specific gene in a specific cell type.

It is easy, like walking on the water

Peter Vesterbacka

4. INFRASTRUCTURE AND SCIENTIFIC PIPELINE DEVELOPMENT

Genetic sequencing has revolutionised the field of biology and medicine by providing detailed information about genes and their functions. The cost of the Human Genome Project was approximately USD 3 billion [6], and sequencing services were expensive. However, the cost of sequencing one human genome has significantly decreased over time. For instance, in 2007, it was estimated to cost around $1 million, while in 2014, it dropped to $1000 [161]. Nowadays, the cost of obtaining a human genome sequence is approximately $200-600 [162,163]. It has been speculated that some of the large DNA sequencing technology companies are striving to reduce the cost of obtaining a genome sequence to as low as $100 per genome [164]. This downward trend in cost has led to a significant increase in the volume of data generated, effectively placing the field of genomics among the "big data" domains [165].

The processing of large datasets, especially in the realm of genomics research, can be a complex undertaking, as personal data are often confidential and the research process must adhere to strict guidelines on reproducibility, traceability and reliability. To address these challenges, scientific software pipelines have become crucial tools for processing, analysing and interpreting big data [166]. It is worth noting that in literature, the terms "pipeline" and "workflow" are often used interchangeably. However, in this thesis, I will primarily use the term "pipeline" for consistency, referring to scientific software pipelines. In the following section, we will offer an overview of the significance of pipelines in scientific research.

To comprehend the significance of pipelines in data analysis, it is essential to take a holistic view of the data analysis process. The process consists of three key elements: the **data** itself, the **software** that outlines how to process the data, and the **computational infrastructure** (i.e. computation power) required to carry out the software instructions on the given data (Figure 15). A pipeline essentially represents a software part of these three key elements. A modern pipeline ought to possess self-sufficiency when it comes to software (i.e. dependency isolation) and be capable of functioning on various computational infrastructures, which could encompass computational clouds, High Performance Computing Clusters (HPC) or even individual computers (i.e. portability).

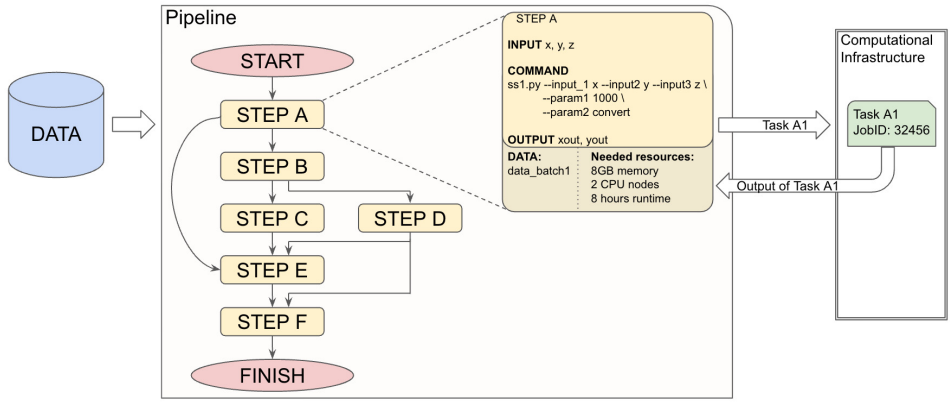

Figure 15: Three key elements to run a scientific analysis with modern pipelines. Data, pipeline and computational infrastructure.

To clarify the typical execution process of pipelines, it is helpful to introduce certain key elements of the system. A pipeline comprises several steps, where each step takes a set of data, applies a software program with specific parameters to the data and sends the resulting output (new data) to the subsequent step(s). Hereafter, each step will be referred to as a “process”, as it is responsible for processing the given data, and the software program with predetermined parameters for processing the data will be referred to as a “command". Essentially, a process is a blueprint for a pipeline step that outlines how to process the given data using the command and what output to expect to pass on to the next process. Once the data are available for processing, a new instance of the blueprint (process) is created with the data attached to it. This newly generated instance is referred to as a “task". The task is then submitted to a job scheduler (i.e. computational infrastructure) to be executed, where it becomes a job. The job waits until the necessary computational resources become available and are reserved for it. When the job execution is completed, the process accepts the expected output and passes it to the next process in the pipeline. Upon reaching the last process in the pipeline, the pipeline execution is considered complete for that particular data batch. Computational infrastructure is generally made up of numerous computational nodes, each with its own maximum capacity for computation (such as Central Processing Unit (CPU) and memory). Although the capacity of nodes in modern infrastructure can be significantly higher than that of typical personal computers, with hundreds of CPUs and Terabytes of memory, it is important to remember that they are always limited.

It is worth noting that certain elements of this chapter have been previously addressed in my master's thesis [151]. Although the developed pipelines have since been refined, the core concepts and features in the pipeline development process have largely remained the same. The following subchapters provide a more in-depth explanation of these features, drawing from the viewpoint of a more experienced developer and incorporating updated literature reviews.

4.1. History of scientific pipelines

Before the advent of pipeline frameworks, researchers typically relied on simple bash scripts or manual data processing methods. However, these approaches posed several challenges. For example, the use of bash scripts often made it difficult to share and reuse code, making it challenging for researchers to build on each other's work. Moreover, these methods frequently lacked the ability to perform parallel processing and were not easily scalable for large datasets. Additionally, these approaches often had issues with reproducibility, portability and reusability, as data processing steps were not standardised and data analysis was often performed in an ad-hoc manner.

As the volume of scientific data grew, it became increasingly clear that a more robust approach was needed. This led to the development of scientific pipeline frameworks, which provide a standardised approach to data processing and analysis, enabling reproducibility and sharing of code. Moreover, these frameworks often have built-in support for parallel processing and scalability, enabling researchers to process large datasets more quickly and efficiently [167]. Additionally, these frameworks usually include tools for data visualisation and quality control [166], enabling researchers to identify and troubleshoot issues in their data. Overall, the development of scientific pipeline frameworks has significantly improved the quality, speed and reproducibility of scientific research [168].

In the field of bioinformatics, there are several pipeline frameworks available with differences in design philosophies, technical issues and ease of use. To gain a better understanding of these frameworks, a review study classified existing pipelines based on three criteria: syntax, paradigm, and interaction [169]. The syntax of explicit frameworks follows the idea of tightly linking tasks together in a certain order, while implicit syntax frameworks calculate intermediate steps automatically. Implicit frameworks are descendants of Make [170], which was developed in the 1970s as an early domain-specific language (DSL). The design paradigm covers principles of frameworks bound to an existing code library instead of independent executables. The interaction paradigm involves command-line-based pipeline frameworks and workbenches that provide users with a graphical user interface (GUI). The most popular workbenches include Galaxy [171], Taverna [172], and commercial cloud-based software as a service (SaaS) workbenches, such as Illumina’s BaseSpace, SevenBridges [173], and DNANexus [174].

Each pipeline framework has its own benefits and drawbacks, and there is no strict formula for preferring one over another. A suitable framework can be selected based on individual preferences, such as coding skills, familiarity with the interface, and the size and complexity of the data being processed [175]. Ultimately, the choice of pipeline framework should depend on the specific needs of the user and the project at hand.

4.2. Computational infrastructure

Given the nature of genomics data, which is typically large in size [165], the analysis of such data requires a significant amount of computational power to perform the necessary computations. Therefore, we require computing resources that are capable of handling large data sets generated by genomics experiments and capable of processing them in a reasonable amount of time and at a reasonable cost. These computing resources should be equipped with a high number of processor cores, ample memory and fast storage devices, such as solid-state drives (SSDs) and high-speed networks, to ensure that the data can be processed efficiently. Additionally, these resources should be scalable to meet the growing demands of the genomics research community as the volume of data being generated continues to increase [167]. This scalability is usually achieved by parallelising the execution of tasks, meaning that instead of executing one task which requires an unrealistic amount of computational resources, the task is split into many smaller tasks and executed in parallel. High Performance Computing (HPC) clusters are used to run these parallel operations [176] and are designed to increase the speed and efficiency of scientific analyses. These clusters typically run on Linux-based operating systems and utilise job scheduling systems, also known as executors, to distribute and manage jobs across computing nodes.

Cloud batch computing services have become increasingly popular because they eliminate the need for users to maintain the cluster, as the service providers take care of it. Additionally, the cost of storing large reference files used in bioinformatics analysis can be high, but some cloud services offer common storage for these files, allowing anyone to use them without incurring additional costs. Cloud services also provide private storage options to keep the raw data close to the computing power, which can reduce costs associated with transferring files. Moreover, some large-scale genomics projects have made their data available on the cloud, which eliminates the need for analysts to download large amounts of raw data to the local computer. Additionally, these services often offer different pricing options, including a “pay-as-you-go” approach, which means that users only pay for the services they use.

There are also some disadvantages to cloud systems that discourage certain institutions from using them. The two main reasons are cost and data security concerns. Firstly, if an institution has a continuous demand for computational power, it is often more cost-effective in the long run to invest in an in-house HPC system [177]. Secondly, even though cloud service providers guarantee the safety of users' data, some research organisations prefer not to use these services due to potential security issues [178].

4.3. Modern scientific pipeline development

In order to handle genomic data, scientists typically guide files through a sequence of defined steps known as a pipeline. For modern research projects to meet the expectations of the research community, these pipelines must possess certain features [179], which have been shaped gradually over time [166,180]. In the following subchapters, we will discuss these features in detail. While I have divided these features into distinct subtopics for the sake of clarity, many of them are interconnected, where inadequate support for one feature could significantly impact the proper functioning of the others.

4.3.1. Reproducibility (replicability)

Modern scientific pipelines are expected to support reproducibility, which means that processing the same dataset with the same set of parameters should produce consistent results regardless of location and computational environment [181,182]. This is necessary for verifying the results presented in publications, which is why major scientific journals now require researchers to publish their data and code [183,184]. Reproducibility is often discussed in terms of "provenance" [185,186], which refers to the origins of input data, tools, results and intermediates in the pipeline. Replicating an experiment in a location and operator-agnostic manner is essential for modern science, particularly in the field of bioinformatics, where data accessibility and availability of tools and guidelines are crucial for reproducing study results [187]. Some studies also attempted to come up with rules of thumb to communicate ways for keeping the research reproducible [188] and promoting a reproducible research culture [189].

The careful handling of genomic data and privacy is critical as it can have severe consequences if misused. The field of genomics is relatively new and technological advancements have led to more frequent discoveries. To take into account the potential power of genomic data, data are usually categorised into two levels of access based on the donor's permission: open access data that is publicly available; and managed access data that requires special permission for use. Therefore, if an experiment involves managed access data, the corresponding permissions must be obtained to reproduce the experiment.

When attempting to reproduce an experiment, if the required data are already available, the biggest challenge is typically related to the technical aspects of the computational tools used. Typically, two common problems can arise:

1. *The software used in the original experiment may not be available or cannot be installed.* A study of bioinformatics software published between 2000 and 2017 found that 26% of tools needed for replication were not accessible, 24% of accessible software failed to install, and 49% were considered difficult to install [190]. The study also found that

publications with accessible and easy installation processes were more highly cited.

2. *Guidelines for using the software may not be available.* In bioinformatics, command-line tools are commonly used, which may generate different results based on the parameters provided. Therefore, information on how the software was used in the original experiment should be provided for consistent results. Additionally, if there are multiple versions of the software, the version used in the experiment should be specified to ensure consistency.

Thus, ensuring the reproducibility of a pipeline is crucial because it consists of multiple steps, where each step requires a specific tool to process the input data. It is important that both the data and tools used in the pipeline are accessible and understandable, with clear explanations provided, if necessary. If any of the tools used in the pipeline are unavailable or produce inconsistent results, the entire pipeline may become useless [191].

4.3.2. Portability and reusability

The field of genetics and genomics data analysis is characterised by its dynamic nature, where novel analytical approaches are consistently developed, affording the possibility of uncovering novel insights from previously collected data. In other words, genetic data that were originally produced for a specific research question can often be utilised for another study that was not initially intended [192]. However, data is the main resource in data science fields, and it is not uncommon for some organisations to withhold it from sharing due to its tangible value. In genomics, there are additional legal obligations to protect the privacy of research participants, which can prevent authors and institutions from sharing the genetic data available to them [193,194]. Overcoming legal barriers to access raw genetic data is often challenging, if not impossible, which highlights the importance of the portability and reusability of data analysis pipelines. For example, local data protection laws may prohibit the transfer of genomic data across national borders. In such cases, one viable solution is to rent private cloud computing resources within the same country and execute the analysis with a portable pipeline. Even though sharing raw data may not be permitted, summary-level analysis results generated from the raw data can often be shared or even publicly released, as the confidentiality of donor information is maintained. Another option is to share the pipeline with the researcher or organisation that has the necessary data access rights, enabling them to conduct the analysis in their own computational environment and share the summary-level data, which is not restricted by legal requirements. Additionally, the portability feature of pipelines offers benefits beyond the ability to use them in physically different locations. In situations where an organisation's infrastructure undergoes changes, the ability to use the same pipeline in the new infrastructure, even if it is in the same location, can also be advantageous [195]. Hence, the concept

of portability involves the capability of pipelines to operate in diverse computational environments, including a range of HPC job scheduler systems (such as SLURM [196], LSF [197], Moab [198], SGE [199], OAR [200], Flux Framework Executor [201], among others), cloud computing services [202] (e.g. AWS, Google Cloud, Azure Batch), and even local machines.

While portability of pipelines is important, it alone is not sufficient to make them useful. Researchers still need to install the pipeline in their computational environment, plug in their data and run the pipeline. To ensure that pipelines are easily usable for end-users, they should be designed for minimal configuration changes to run successfully. This is known as the reusability of the pipeline, which is measured by how easy it is for users to install and execute the pipeline on their own data. The pipeline should be user-friendly enough that even users without prior knowledge of the field can operate it with minimal effort. Additionally, pipeline flexibility, such as the ability to change parameters, can increase its reusability by better fitting it to the user's data and execution environment. Although there is currently no standard test to measure the reusability of a bioinformatics pipeline, feedback from users and popularity within the community are considered to be key indicators [166].

4.3.3. Parallelisation and scalability

To execute each step of a pipeline, specific hardware resources such as memory, time, and CPU cores are required to process the input data and pass the output data as input for another step. The resource requirements of the tasks are typically directly proportional to the size of the input file. When the input file size is too large, the computational environment may lack the computational node powerful enough to handle execution of the task. For these cases, the software used in a step may provide options for dividing the input into smaller chunks and processing them in parallel as multiple tasks, using fewer resources per task [203]. Thus, a software program is considered scalable if the pipeline does not crash due to the size of the input data and successfully completes the execution of all tasks [195]. Since a pipeline consists of tools configured in a specific manner, all the tools within the pipeline must also be scalable for the pipeline to be considered scalable. Ideally, if (theoretically) there is an unlimited amount of available computational power, the execution time of the pipeline should not change according to the size or volume of the input data.

4.3.4. Dependency isolation

In data science, it is common practice to use existing tools and libraries. This process often leads to the development of software that is dependent on other software. Additionally, if any of the dependencies have specific requirements such as a particular operating system or environment, the dependent software also inherits these requirements. In the context of bioinformatics pipelines, the situation is very similar, especially when pipelines are script wrappers that use multiple external tools to process and analyse data in a structured manner.

Certain programming languages have techniques to isolate program dependencies from those of the operating system. This can be achieved, for instance, through virtualenv for Python, Carton for Perl, or Bundler for Ruby. While these tools are effective in creating and maintaining independent environments for single-language programs, other solutions are available for managing environments that involve multiple programs developed in different languages. One such tool is Conda [204,205], which can create virtual environments with a given "recipe" containing the required software and installing all the tools in the recipe with a single command. Virtual machines like VirtualBox [206] and VMWare [207] are also available, especially when a graphical user interface for a virtual environment is necessary. However, they are not commonly used in research projects that involve vast amounts of data.

Currently, the most popular way to achieve dependency isolation is through the use of software containers [208]. Unlike Conda, which only saves the recipe of the needed tools to perform the computation, containers save the tools themselves to run the pipeline. Containers have many advantages over other dependency isolation methods. First, they are highly portable, enabling easy access to dependencies and in turn ensuring portability and reproducibility of the pipeline [209]. Custom scripts can also be added to a container without limiting the portability of either the container or the pipeline. For openness and tractability, it is important to provide the exact content of the container, including the tools with their corresponding versions. Docker [210] and Singularity [211] are the most popular software container tools currently used. Base images, which typically contain basic container needs such as an operating system, are used as a baseline environment in which to build the needed tools into the containers. A recipe file, such as Dockerfile, is used to provide step-by-step instructions for how to build the container. Containers can be stored in web repositories, such as Docker Hub, Singularity Hub, or quay.io, where they can be easily shared and pulled. Docker, which is designed for enterprise software production systems, gives superuser privileges to the user. In multi-user systems like HPCs, this can create problems, since regular users do not have administrator privileges [211]. Singularity, which behaves like Docker, can be used without administrative privileges. Finally, containerisation makes it relatively easier to automate the testing of dependencies [212], which is important for continuous integration and delivery in software development [213].

4.3.5. Adopted practices from software engineering

In addition to the above-mentioned techniques used in pipeline development, there are practices that have been adopted from software engineering, including but not limited to testing, continuous integration (CI), versioning, modularity, and extensibility. The adoption of these practices serves to improve the reliability and maintainability of the pipeline.

Testing and Continuous Integration. Modern pipelines are often open-source and community-driven, allowing anyone to examine and verify the

source code to ensure that the software functions as intended. Typically, a small group of developers initiates the development of the pipeline, and after the initial release, users of the software may suggest or contribute new features to the project. To prevent chaos, the pipeline usually has a few maintainers who make decisions about the acceptance of the changes and contributions offered. With contributions open to anyone, it is beneficial to have automatic checks in place that verify changes before they are reviewed by the maintainers. Hence, automated testing [214] and CI are critical for open-source [215], community-driven pipeline development in order to ensure that changes do not create issues in the currently working version of the pipeline [166].

Software versioning is the process of assigning a unique version number or name (i.e. tag) to a particular release of software. This is an important practice in software development, including pipeline development. Versioning helps developers to keep track of the changes made to the software and to communicate these changes to other developers and users. Additionally, in order to ensure replicability of the results, it is essential to utilise the accurate version of the pipeline. In short, software versioning provides benefits such as improved collaboration and smoother software releases.

Modularity and Extensibility. The extensibility of a pipeline refers to its ability to be easily extended or utilised as a foundation for the development of new pipelines. In scientific research, it is often necessary for researchers to build on existing pipelines in order to create new ones. Modularisation of pipeline steps has become a popular approach in many institutions, making it easier to reuse these steps in other studies. The pipeline's steps can be compared to Lego blocks, or modules, which can be combined in different ways to create a new pipeline. Additionally, the modularity approach enables checkpointing, where re-running the failed pipeline execution does not re-run successfully completed steps but continues the execution from the failed step [167,195]. These concepts of extensibility and modularity align well with the Open Closed Principle [216] widely used in software engineering. According to this principle, a software component should be designed in such a manner that it can be easily extended with new functionality, without needing to modify its existing source code. In essence, the component remains “open” for the incorporation of new features, while remaining “closed” to any alterations in its existing behaviour. This approach aims to enhance the robustness and adaptability of software systems, enabling them to evolve over time while minimising the risks associated with code changes.

4.4. Challenges in scaling and reusability/portability (federated analysis)

Despite efforts to address scalability in modern pipelines, there are still obstacles that need to be overcome. Firstly, scientific software pipelines are built to handle inputs of any size. However, in practice, this may not always be feasible. Sometimes the input size can be so large that the process command may theoretically be able to complete the execution, but it would take an impractical amount of time. In such cases, more specialised tools can be used to convert the data into a smaller, more representational format [217], which can then be processed by tools that accept the newly generated format. Secondly, sometimes the infrastructure the pipeline is using for execution can have hardware limitations that cannot be addressed in pipeline design. For example, if one process in the pipeline has to make too many Input-Output (IO) operations in a short time, it may burden the file system of the corresponding infrastructure. There are certain workarounds to address these kinds of issues, but these workarounds are usually infrastructure-dependent. Hence, it has become rather straightforward to use publicly available pipelines with minimal effort only if the data size is not unexpectedly large. For edge cases, however, some experience in pipeline development is usually required to overcome atypical challenges.

Another challenge lies in the portability and reusability of the pipeline. The reusability of pipelines becomes a significant concern, especially in federated analysis where the pipeline is shared with researchers in different locations for data analysis. Although pipelines are designed to be adaptable to various infrastructures (i.e. portable) and come with documented instructions and example input for testing purposes, individuals who are not familiar with the pipeline may still find it challenging to run.

Talent wins games, but teamwork and intelligence win championships

Michael Jordan

5. EQTL CATALOGUE (PUBLICATION I)

Researchers persistently explore various methods and innovative approaches to decipher the explanations for GWAS signals [218]. GTEx, for instance, is a large-scale study that aims to understand how genetic variation influences gene expression across various human tissues, providing valuable insights into gene regulation and its relationship with disease [51]. The latest version of this resource (i.e. Version 8) includes genetic data from 838 postmortem donors and 15,201 RNA-seq samples across 49 tissue sites [219]. eQTLGen, on the other hand, is a collaborative initiative focusing on the identification and analysis of eQTLs, which includes 31,684 blood samples from 37 consortium cohorts, and aims to better understand the genetic underpinnings of complex traits and diseases [48]. Extensive research by GTEx and eQTLGen consortia has shown that detecting many regulatory effects in bulk tissues under steady-state conditions is difficult. However, studying disease-relevant cell types and tissues has been successful in finding additional colocalisation signals not detected in the GTEx study. A possible solution is to combine various context-specific QTL study results into one database to make downstream analysis easier.

Several efforts have been made to create a QTL database by compiling original results from different QTL studies. However, most datasets in these databases are incomplete, missing critical information such as effect alleles, standard errors, and sample sizes, which are all important for downstream analyses like colocalisation and Mendelian randomisation. What is more, the technical variation between studies in sample collection, RNA-sequencing protocols, genotyping, and data analysis raises questions about the influence of technical differences on eQTL effect sizes and the sharing of eQTLs between cell or tissue types. Analyses based on GTEx data have estimated high levels of eQTL sharing between most bulk tissues, while smaller studies have often estimated much lower levels of sharing between purified cell types. However, these analyses are sensitive to how sharing is defined, which genes and variants are included, and which analytical approaches are used, making it impossible to directly compare eQTL sharing estimates between studies.

To overcome these issues, a potentially better approach to building this database is to collect individual-level data and uniformly process all QTL studies, but this is challenging due to limitations on sharing individual-level data. In the eQTL Catalogue project, we chose a more challenging approach by gathering individual-level data from 21 separate studies and processing them uniformly. The results showed that eQTL effect sizes from matched cell types or tissues were highly reproducible between studies, and differences in eQTL effect sizes between datasets were primarily due to biological differences rather than technical differences in sample processing. Uniformly processed summary statistics enabled the characterisation of eQTL diversity across 69 distinct cell types and tissues, revealing high levels of cis-eQTL sharing between most bulk tissues but a much smaller proportion of eQTLs shared between purified cell

types. This eQTL diversity also manifests in disease colocalisation, detecting many novel colocalisations that are missed when analysing GTEx data alone. The uniformly processed QTL summary statistics and fine-mapping results are available from the eQTL Catalogue FTP server and REST API as well as through the Ensembl Genome Browser. It is possible to navigate to these resources through the eQTL Catalogue website (https://www.ebi.ac.uk/eqtl/). Additionally, the summary statistics are integrated into Open Targets Genetics portal and can be explored freely.

My contribution to this resource was significant. Initially, I took the lead in developing the three pipelines for RNA-seq quantification [116], quality control and normalisation [117], as well as QTL mapping [118]. Drawing on my prior experience in software engineering, I applied best practices for pipeline development, which are detailed in Chapter 4. The most important decisions made in the pipeline development of the eQTL Catalogue project are outlined in Table 2. Firstly, we decided not to develop some pipelines from scratch but to extend already available pipelines. For example, we adopted the RNA-seq pipeline from nf-core [166] and added the quantification methods we desired for the project. We tried our best to keep the pipelines under development reusable by other researchers and developers. Secondly, we made sure that the pipelines support main features such as parallelisation, scalability, dependency isolation, modularity, and extensibility. Lastly, we provided a small set of test files to ensure the portability, testability and easy reusability of the pipelines. Eventually, these pipelines happened to be very impactful (i.e. QTLmap pipeline GitHub repository has 33 stars), and, to the best of our knowledge, are currently being actively used by labs at the Wellcome Sanger Institute, the Finnish Institute of Molecular Medicine, the Estonian Biobank, Helmholtz Munich and likely many other places we are not even aware of. Subsequently, I utilised these pipelines and my data analysis expertise to process the data and conduct downstream analyses on the resulting QTL summary statistics. In the end, the culmination of these efforts resulted in a publication, co-authored by myself, incorporating contributions from all authors involved.

Table 2: Some of the decisions made in the pipeline development of the eQTL Catalogue project

#	Decision	Outcome	Affected principle(s)
1	Develop separate pipelines for each step of eQTL Catalogue workflow, instead of developing a single gigantic workflow	The pipelines do not depend on each other. For example, change on *qcnorm* pipeline does not require re-running all datasets through the *rnaseq* pipeline.	Modularity, portability, reusability
2	Use Nextflow Domain Specific Language (DSL) and workflow engine	Nextflow inherently takes care of portability between task schedulers and parallelisation and supports modularity. Hence, adopting Nextflow, despite being in its early ages, made our pipelines more scalable, portable, modular and extensible.	Portability, scalability, parallelisation, modularity, extensibility
3.	Use software containers	Users do not need to install any tools used in the processes of the pipelines, except the very few core programs (i.e. Nextflow for pipeline execution, Singularity or Docker for container usage)	Dependency isolation, reusability, portability, reproducibility, modularity
4.	Containerise the used software per process and not per pipeline	Containerised processes (i.e. individual steps within pipelines) allows for their integration into different pipelines; updating a process necessitates modifying only a single, smaller container rather than updating large containers (as would be the case with pipeline-level containerisation).	Dependency isolation, modularity
5.	Version all pipelines and containers	All pipelines and the containers within them are versioned, ensuring that a particular GitHub release or tag is consistently linked to the specific versions of the relevant containers.	Dependency isolation, modularity, extensibility, software versioning
6.	Provide a small set of test data	All three core pipelines (i.e. *rnaseq*, *qcnorm* and *qtlmap*) contain a small set of test data. This additional data enhances reusability for other users and provides an extra layer of validation when updating the pipeline.	Testing, reusability, portability

It is crucial to acknowledge that collecting, processing, and analysing such a vast amount of data is an incredibly challenging task, if not impossible, for a single individual. In the eQTL Catalogue project, for instance, data collection and genotype imputation pipeline development were primarily carried out by Kaur Alasoo, while James Hajhurst almost single-handedly developed and implemented the eQTL Catalogue API. Liis Kolberg was responsible for microarray gene expression data normalisation and quality control, while Kateryna Peikova conducted eQTL similarity and matrix factorisation analyses. Other co-authors contributed based on their areas of expertise. This invaluable resource would not be available without their combined efforts, and I am deeply grateful to have collaborated with such a talented team.

5.1. Uniform processing and Quality Control (QC)

To process a wide range of eQTL studies uniformly, we developed a robust and modular data pipeline. First, we carried out thorough QC and imputed the genotypes. For RNA-seq datasets, we conducted QTL mapping for four molecular traits (i.e. gene and exon expressions, transcript and transcriptional event usage; see chapters 3.1.2.1–3.1.2.4), performing the analysis individually for each dataset, such as cell type or tissue, within each study. We found the highest number of QTLs at the gene expression level, and the number of significant associations increased linearly with sample size. We noticed a similar linear trend for microarray datasets, but downstream analyses focused mainly on RNA-seq-based eQTL datasets due to their wider range of cell types and tissues as well as on making up most of the samples in the eQTL Catalogue.

5.2. Novel colocalisations relative to GTEx

We performed a colocalisation analysis between GWAS summary statistics for 14 traits and either the new eQTL Catalogue datasets or all GTEx tissues. This analysis led us to discover that the eQTL Catalogue contains many eQTLs not present in GTEx, which could potentially improve the interpretation of complex trait and disease associations. We identified at least one colocalising eQTL for 4,528 independent loci across 14 traits, with 20.4% not captured by GTEx. The additional colocalising loci ranged from 14% for height to 29% for lupus, indicating that relying solely on GTEx could lead to missed trait colocalisations. Moreover, we observed additional colocalisations even in eQTL Catalogue cell types and tissues already captured by GTEx, which could be attributed to thresholding effects, increased sample sizes, or biological and population differences between datasets.

5.3. Transcript-level quantification methods finds additional colocalisation signals

We also conducted a colocalisation analysis across 14 complex traits and all three transcript-level QTLs and discovered that 16.7% of colocalisations in independent LD blocks [220] were detected using only one of the three transcript-level traits, and not by traditional eQTLs in any of the 95 RNA-seq datasets. Even so, the percentage of additional signals identified might be understated, as transcript and gene-level QTLs could colocalise with distinct GWAS variants within the same LD block. If any of the variants in the independent LD block colocalised with eQTL, they were not counted as additional signals.

5.4. Invisible efforts and lessons learned

After years of researching and developing software products for data analysis, one thing I learned is that the data analysis should be reproducible by others and extremely easily reproducible by the person who did it in the first place. When conditions and ideas change, ideally, the corresponding analysis modifications should take significantly less time than doing the analysis from scratch.

Another essential lesson learned is the need for manual quality control. While we have designed pipelines to automate QC to some extent, the pipeline essentially generates data for human inspection of data quality. Ultimately, a human being manually examines the produced QC materials and determines the appropriate course of action for potentially problematic samples. Typically, there are three possible outcomes: exclude the sample, retain the sample, or, if feasible, correct the problematic sample. Although this may appear to be a somewhat easy task, it requires a solid understanding of the sample context and general molecular biology, particularly for samples that fall within the borderline "problematic" range.

Finally, published and recommended best practices may be not the best ones for the project. When trying to utilise the Leafcutter for splice-junction usage quantification, WASP [221] was the recommended tool for reference mapping bias adjustments. It is also an embedded feature of STAR aligner, which made us consider switching from HISAT2 to STAR for mapping the reads. Nevertheless, after preliminary testing, we discovered that WASP is not appropriate for our objectives, as it eliminated a large proportion of reads originating in gene-coding areas, leading to a significant reduction in statistical power and an inability to identify widely recognised QTLs. Furthermore, WASP can only utilise SNPs, excluding insertions and deletions. In light of these findings, we continued to use HISAT2 as the alignment tool in the RNA-seq quantification pipeline and, instead, relied on visualisation approaches to detect false positives caused by reference mapping bias.

Use a picture. It's worth a thousand words

Arthur Brisbane

6. SYSTEMATIC VISUALISATION OF MOLECULAR QTLS (PUBLICATION II)

The majority of genetic variants linked to complex traits are found in non-coding regions of the genome. MolQTL studies over the past decade have shown that genetic variants in non-coding regions regulate gene expression levels, splicing, promoter usage, or alternative polyadenylation. Although the eQTL Catalogue has contained transcript-level QTL summary statistics since its first release, pinpointing the precise mechanism of action for each molecular QTL has remained challenging due to overlapping QTLs detected by different RNA-seq quantification methods, technical biases in read alignment, and the many alternative transcripts or splice junctions to be considered for each gene. Additionally, determining the size and direction of genetic effects, the affected gene parts, and the absolute expression of impacted transcripts is often difficult based on summary statistics alone because the usage of each transcript or splice junction is quantified relative to all other transcripts of a gene.

The challenge can be addressed by visualising the average RNA-seq read coverage stratified by each possible genotype of a QTL variant. QTL coverage plots have been used to characterise chromatin QTLs and to verify promoter usage and splicing QTLs. However, systematically applying this approach to large molecular QTL collections like the GTEx project and the eQTL Catalogue is challenging due to the separate stratification needed for each significant genetic variant and molecular trait pair, making the number of QTL coverage plots required potentially unmanageable.

In this publication we developed a novel approach for generating QTL coverage plots for all independent genetic signals and related molecular traits. This involves updating data processing workflows to enhance promoter usage and splicing QTL discovery, generating read coverage files (i.e. BigWig) for 25,724 RNA-seq samples, and adopting fine-mapping-based filtering to identify all independent genetic signals and molecular traits while reducing the file size of summary statistics by 98%. To accommodate colocalisation methods accounting for multiple independent causal variants, signal-level log Bayes factors were computed for all independent signals. This approach allows the definition of tag variants for all independent genetic associations across 127 eQTL datasets and the creation of QTL coverage plots for more than 1.7 million QTLs for interpreting most of the colocalising signals within the eQTL Catalogue.

In the following sections, I will discuss some key elements, including filtering through fine-mapping, visualising coverage plots, and the difficulties we encountered in developing the relevant pipeline. However, I want to acknowledge that this publication is not solely my own work. Especially, I would like to recognise the contributions made by Ralf Tambets, who modified the QTL mapping workflow to support fine-mapping and performed colocalisation analysis against vitamin D GWAS, as well as Kaur Alasoo and Peep Kolberg, who made multiple extensions to the genotype imputation pipeline.

6.1. Fine-mapping-based filtering

A significant challenge we encounter with exon-level and transcript-level associations is the large number of correlated traits tested, resulting in large summary statistics files. For example, we have 521,889 distinct exons, 215,956 distinct transcripts and 392,588 transcriptional events from a pool of 38,114 possible genes, of which 19,964 are protein coding. When combined with the count of genetic variants in *-cis*, the number of tested associations easily reaches the scale of billions, which results in large files. For instance, even after compression, exon QTL summary statistics for one dataset are usually larger than 40 Gigabytes. To address this, we implemented fine-mapping-based filtering, using fine-mapped credible sets to identify all independent signals at the group level (e.g. gene for transcript usage; leafcutter cluster for splice-junction usage; trait group for transcriptional event usage) and retaining only the most strongly associated molecular trait for each signal. Subsequently, we identified overlapping variants between credible sets and defined these as connected components (Figure 16). For each connected component, we chose the molecular trait with the highest posterior inclusion probability (PIP) and retained only the summary statistics of these selected molecular traits.

Figure 17A provides a demonstrative example featuring the *CYP2R1* gene, which encompasses five transcriptional events—hereafter referred to as molecular traits—within the group designated as *ENSG00000186104.grp_1. contained.* These molecular traits exhibit a relative usage pattern; an increase in the usage of one trait inversely affects the usage of others, as depicted in Figure 17B. To choose the molecular trait with the strongest association, first we filter out credible sets with over 200 variants or a z-score below 3. In the presented example, three out of the five molecular traits—specifically ENST00000525015, ENST00000532378, and ENST00000532805—did not possess any credible sets that met these established criteria. Consequently, the two credible sets belonging to the remaining two molecular traits—ENST00000334636 and ENST00000534686— form a connected component (Figure 16, Figure 18A). In the given example, the remaining credible sets shared a variant, forming a connected component. The lead signal of this connected component was from the credible set of ENST00000334636, as it had the highest PIP value, hence this molecular trait is selected as a representative of the group (Figure 18B). To be clear, after initial filtering, if the remaining credible sets do not share variants, they each become an independent connected component. Then, we choose the signal with the highest PIP from each connected component to represent the group. This means there might be multiple signals acting as representatives for the group.

This approach reduces the size of summary statistics files by around 98% while preserving nearly all significant associations for colocalisation purposes. Additionally, the reduction in univariate summary statistics file size enables us to export SuSiE log Bayes factors for each fine-mapped signal and tested

variant, which can be directly utilised in the new coloc.susie method to perform colocalisation analyses between all pairs of independent signals [67].

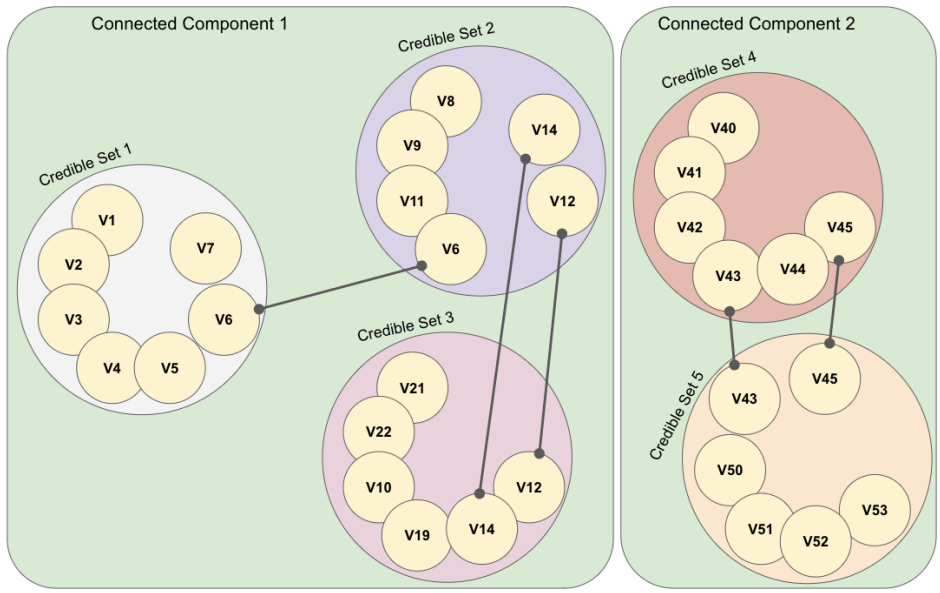

Figure 16: Building connected components from credible sets. Each credible set contains one or more genetic variants. If two credible sets share a variant they form a link between each other and a cluster of linked credible sets forms a connected component.

6.2. Leafcutter sQTLs + visualisation

The previous release of the eQTL Catalogue included four molecular trait quantification methods to measure transcriptional changes from RNA-seq data: gene expression (ge), exon expression (exon), transcript usage (tx), and transcriptional event usage (txrevise). In addition to these four, we have now implemented LeafCutter [150] to directly quantify the usage of splice junctions (Figure 11). By utilising fine-mapping-based filtering for these four molecular trait QTLs, we now have a manageable collection of independent lead variants and related molecular traits across all datasets. This enables us to generate visualisations of coverage plots that contain normalised RNA-seq read coverage across all exons of the gene (Figure 17A), exon-level QTL effect sizes, and standard errors (Figure 17B) as well as the alternative transcripts or splice junctions employed in association testing (Figure 17C). At present, we have created and shared only static versions of the coverage plots. However, plans are underway to develop a platform that will allow users to engage with dynamic coverage plots in the near future.

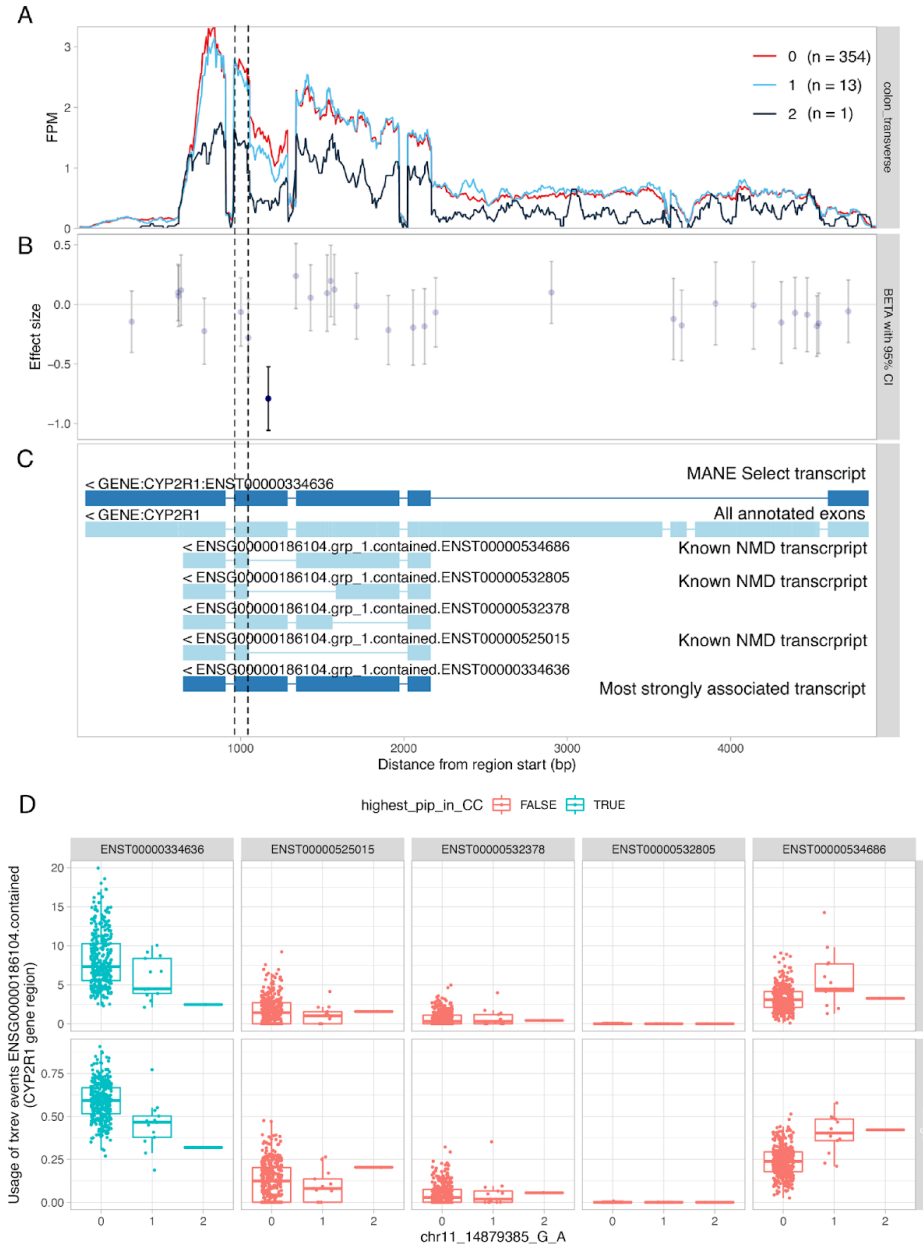

Figure 17: Visualisation of a splicing QTL detected in the *CYP2R1* gene. (A) RNA-seq read coverage across the *CYP2R1* gene in GTEx transverse colon tissue stratified by the genotype of the lead sQTL variant (chr11_14855172_G_A). All introns have been shortened to 50 nucleotides with wiggleplotr [222] to make variation in exonic read coverage easier to see. (B) Effect sizes and 95% confidence intervals of the lead sQTL variant on the expression level of individual exons (or exonic parts) of *CYP2R1*. Associations significant at FDR ≤ 1% are shown in dark blue. (C) The top two rows show the MANE Select [223] reference transcript and all annotated exons of *CYP2R1*,

respectively. The remaining rows show the txrevise [79] event annotations used for sQTL mapping. The short version of exon 4 (between dashed lines) is only present in annotated nonsense-mediated decay (NMD) transcripts. (D) Boxplots of individual txrevise events stratified by the genotype of the lead QTL variant. Top row shows absolute normalised counts in TPM units and bottom row relative usage of each transcript. The selected tag transcript is highlighted in cyan. Figure is taken from supplementary figures of Kerimov et. al. [127] and adapted to the thesis.

6.3. Fantastic challenges and where to find them

To create these visualisations, I developed a new pipeline that takes all required input files and generates coverage plots, box-plots, and essential anonymised data for recreating visualisations in the future (Figure 17). While the pipeline can divide large files into smaller chunks for parallel processing, pipeline execution relies heavily on numerous small input/output (IO) operations. For instance, in the University of Tartu's HPC, where we conducted the analysis, these IO operations placed an additional strain on the cluster network, which in turn slowed down pipeline execution and occasionally caused issues for other HPC users. After exploring various possible solutions, we ultimately reserved several high-capacity computation nodes, copied the IO-intensive files to these nodes before executing the pipeline and ran the pipeline on the reserved nodes. This method removed the strain from the network file system, as the IO operations occurred on the same node where the pipeline was running. Although this approach had its drawbacks, they were easy to resolve and did not necessitate the development of new strategies. Although I have made it sound like an easy problem to solve by summarising in a few sentences, in reality, it took us about six months of trial and error before we found the right solution (Figure 17).

Initially, we planned to share the data necessary for regenerating the coverage plots and designed the pipeline accordingly so that it is also capable of accepting these generated visualisation data. It would be beneficial if other researchers were to generate the coverage plots in their own style or integrate more data to expand the potential of the resource. However, we decided not to share these visualisation data publicly for two reasons. Firstly, the data include some partial individual-level information, such as the genotype of the QTL variant, and although we anonymised the data, we were concerned about potential de-anonymisation attempts. Secondly, it would be an additional resource burden to store and efficiently share this large volume of data. Instead, the data will be used in the development of an interactive coverage plot visualisation tool.

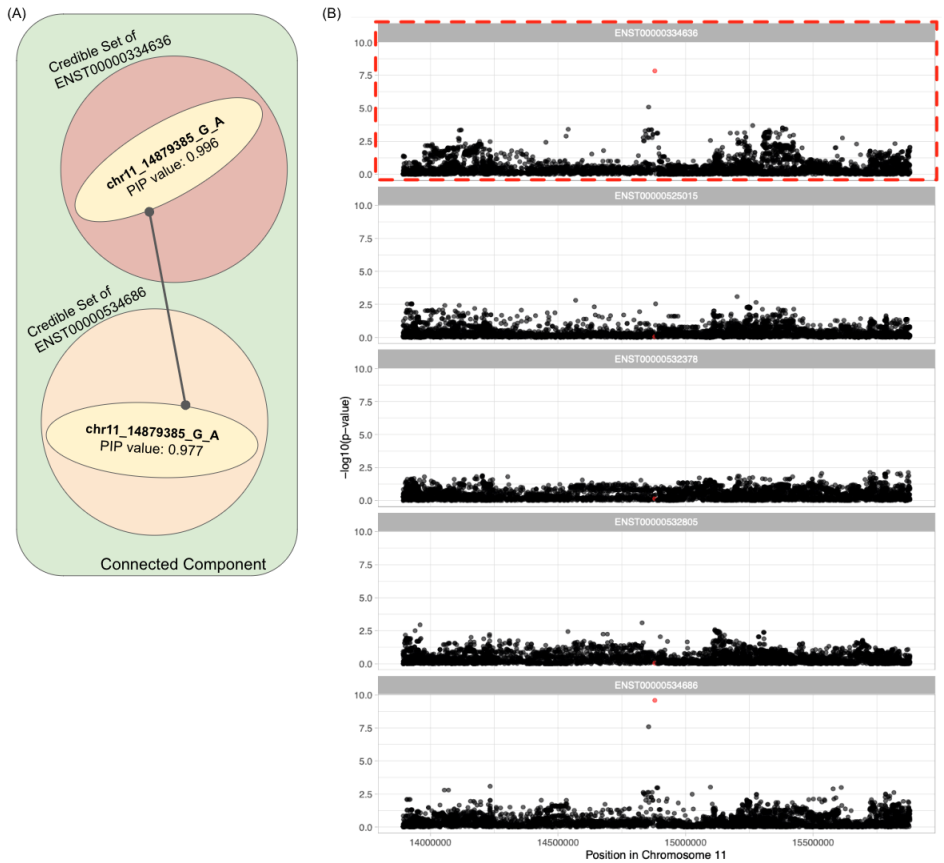

Figure 18: Fine-mapping-based filtering. (A) The connected component of credible sets belonging to *CYP2R1* gene transcriptional events (i.e. molecular traits). Only 2 (ENST00000334636 and ENST00000534686) out of 5 molecular traits have credible sets after initial filtering. (B) Regional association plots for all molecular traits belonging to the *ENSG00000186104.grp_1.contained* group of the *CYP2R1* gene. ENST00000334636 is selected as a molecular trait representing the group (marked with dashed red rectangle). Panel (B) is taken from supplementary figures of Kerimov et. al. [127] and adapted to the thesis.

In the end we retain from our studies only that which we practically apply

Johann Wolfgang Von Goethe

7. APPLICATION OF THE EQTL CATALOGUE TO GWAS INTERPRETATION (PUBLICATIONS III AND IV)

In this section, I will discuss the advantages of having readily available, reusable pipelines and explore how the eQTL Catalogue serves as a valuable resource for other researchers.

7.1. Easy-to-use pipelines facilitate contributions to renowned research projects

A major benefit of the eQTL Catalogue is that both the uniformly processed, formatted and quality-controlled datasets as well as the scalable computational pipelines facilitate rapid reanalysis of data as novel methods become available. For example, the first public release of the eQTL Catalogue in January 2020 did not contain any fine mapping results. However, in discussion with the FinnGen project, it became obvious that providing eQTL fine mapping results could really benefit the interpretation of their GWAS findings. Thus, we implemented SuSiE fine mapping in our QTL analysis pipeline, reanalysed all of the quality-controlled eQTL Catalogue datasets with the updated pipeline and provided the FinnGen project with the fine mapping results. This was only possible because of the significant early investment we had made into obtaining access to individual-level raw data and performing stringent QC on all datasets (Figure 19). Some of the results from this collaboration are described in the recently published research article "FinnGen provides genetic insights from a well-phenotyped isolated population" by Kurki et al. [224] (Publication III). The eQTL Catalogue was used as a valuable resource to investigate potential mechanisms of action for the 27 newly discovered associations. Of the 27 loci, the team identified potential causal coding variants in nine and colocalisation with eQTL Catalogue eQTLs in four. While the disease relevance of these colocalising eQTLs was not apparent in this study, utilising all available tools to help explain the molecular mechanisms underlying diseases remains a valuable approach.

The comparison between genomics research and gold hunting is quite fitting. Gold hunters use tools like metal detectors, pickaxes, and shovels, while genetics researchers have tools such as GWAS, fine-mapping, and colocalisation, with the eQTL Catalogue being one of them. These tools help in searching for the underlying causes of diseases at the molecular level. Although using these tools does not guarantee finding "gold", it does make the exploration process more manageable and efficient.

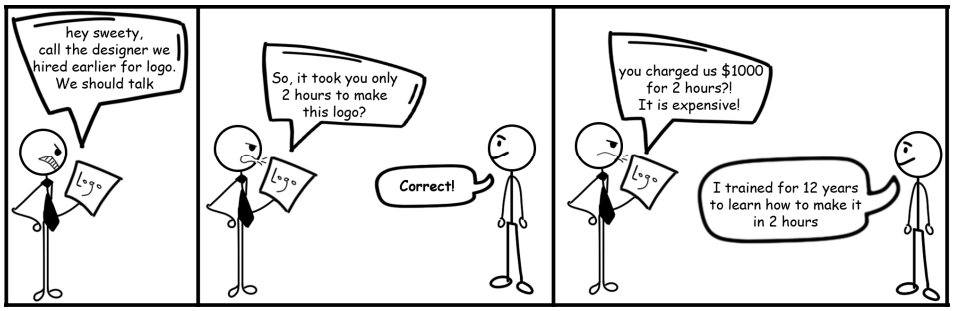

Figure 19: The value of developing easily reusable pipelines and transferable skills. The sketch was initially created by Kamila Mammadova upon my request and finalised by myself.

I also contributed to another study led by Masahiro Kanai ("Insights from complex trait fine-mapping across diverse populations" [15]). In this study, Masahiro performed statistical fine mapping for 117 complex traits across three large biobanks (BioBank Japan, UK Biobank and FinnGen). This study led to the identification of 4,518 variant-trait pairs with a high probability of causality (posterior probability > 0.9). Among these, 285 pairs were replicated across multiple populations. By examining the Finnish and Japanese populations, they discovered 21 and 26 potential causal coding variants, respectively, with significant allele frequency enrichment (i.e. specific genetic variant occurs much more frequently in a particular population than others). By aggregating data from various populations, they identified 1,492 unique fine-mapped coding variants and 176 genes with multiple independent coding variants influencing the same trait. My contribution to this study was performing eQTL analysis and statistical fine mapping on the eQTL Catalogue datasets. This analysis revealed that while fine mapped eQTLs from the GTEx project colocalised with 2331 GWAS credible sets, including the eQTL Catalogue datasets into the analysis increased the number of colocalisations detected to 3787 (62% increase). This study demonstrates that fine-mapping across diverse populations can offer new insights into the biology of complex traits by pinpointing high-confidence causal variants for further analysis.

Another benefit of the pipeline infrastructure and quality-controlled datasets is that they have directly enabled multiple other research projects in the lab that make secondary use of these datasets. For example, Liis Kolberg was able to reuse many of the workflows and quality control procedures to perform large-scale *trans*-eQTL analysis across five datasets (Figure 19) [225].

7.2. Applications of the eQTL Catalogue

In the previous section, I discussed research projects that made use of the eQTL Catalogue in conjunction with our expertise in pipeline development and application. Numerous other researchers have employed the publicly available eQTL Catalogue summary results for various downstream analytical purposes.

For example, GWASs for Migraine [226], otosclerosis [227], Sjögren's syndrome [228], Alzheimer's disease [229], and neonatal jaundice [230] used eQTL Catalogue summary statistics for colocalisation analysis. Other studies used the resource in attempt to understand and prioritise the functional genes and other genetic basis of complex traits, such as pelvic organ prolapse [231], cervical cancer [232], asthma [233], pneumonia [234] and, inflammatory and infectious upper respiratory diseases [235]. Furthermore, numerous studies used the eQTL Catalogue in research not directly studying specific diseases, for example, studies devising methods to systematically prioritise causal variants [70], integrating downstream analysis of other molecular QTL studies [236], and examining various other genetic regulatory activities as well as validating identified signals [237–240]. It is also important to note that eQTL Catalogue data are integrated into various platforms for visual exploration of and connection to other kinds of genomics data, such as Open Targets Genetics [70], Ensembl [241], and FIVEx (an interactive multi-tissue eQTL browser) [242].

Although the studies mentioned above are an obvious measure of usefulness and impact of the eQTL Catalogue in the field, we are dedicated to continuously improving this resource by adding new uniformly processed data and updating our pipelines to adopt cutting-edge tools to extract as much reliable information as possible.

DISCUSSION AND CONCLUSIONS

We have created and openly shared a compendium of uniformly processed human molecular QTLs, encompassing traits such as gene and exon expression, transcript and transcriptional event usage, and splice-junction usage. This resource includes data from 74 distinct cell types and tissues, along with several environmental stimuli. The primary advantage of our eQTL Catalogue is that all datasets are processed uniformly and summary statistics are readily available for subsequent analyses, such as colocalisation. Furthermore, we have developed a wide range of open-source scientific pipelines that can be utilised by other research groups or simply studied to gain a better understanding of how the data are generated.

Following the publication of our QTL summary statistics, we received feedback indicating difficulties in interpreting splicing QTL signals. Hence, we decided to generate QTL coverage plots for all independent genetic signals and related molecular traits. This involved updating data processing pipelines, generating read coverage files for 25,724 RNA-seq samples, adopting fine-mapping-based filtering, and computing signal-level log Bayes factors for all independent signals. This enabled us to identify tag variants for all independent genetic associations across 127 molecular QTL datasets and create QTL coverage plots for over 1.7 million QTLs within the eQTL Catalogue. Our subsequent colocalisation analysis revealed that splicing QTLs are significantly less pleiotropic than eQTLs. Another finding for us was the realisation that generating more than 1.7 million plots is not technically straight-forward task but requires careful design considerations.

The eQTL Catalogue's uniformly processed datasets and computational pipelines enable rapid reanalysis as new methods become available (Publications III and IV). For instance, SuSiE fine mapping was implemented in the QTL analysis pipeline, assisting the FinnGen project in interpreting their GWAS findings. The pipeline infrastructure has also supported numerous other research initiatives, such as Liis Kolberg's extensive *trans*-eQTL analysis across five datasets [225].

The eQTL Catalogue resource has been used in various research projects for different analytical purposes, such as colocalisation analysis in GWASs for various diseases and understanding the genetic basis of complex traits. It has also been integrated into various platforms for visual exploration and connecting other genomics data. The Catalogue has demonstrated its usefulness and impact in the field, and the team is committed to continuously improving this resource by adding new data and updating pipelines to adopt cutting-edge tools for more reliable information extraction.

Despite its considerable impact, the eQTL Catalogue is not without limitations. Obtaining QTLs from all biological contexts is a significant challenge; there are certain cell types we have not been able to include due to legal barriers regarding dataset access. To build a more comprehensive and diverse catalogue

of uniformly processed QTLs, we advocate for the normalisation of federated analysis approaches. In line with this goal, we will continue to enhance the reusability and portability of our data analysis pipelines, making them available through community initiatives like the nf-core [166] repository. Moreover, numerous single-cell RNA-seq eQTL datasets have emerged from studies involving induced pluripotent stem cells (iPSCs) and peripheral blood cells, with more expected to follow. These datasets could dramatically enhance our understanding of cell-type-specific gene regulation in complex tissues. Some of these single-cell RNA-seq studies have since begun to be incorporated into the eQTL Catalogue by other team members and will be made available upon completion.

One last challenge pertains to our coverage plots. Currently, these visualisations for interpreting QTLs are static, and providing interactive versions of these visualisations would enhance their utility. I firmly believe that developing such a tool would significantly aid explorative analysis. Despite the challenges, our team remains committed to enhancing the eQTL Catalogue and its usability, continuing to support the advancement of genomic research.

In my view, it is crucial to consider this resource as "an additional tool in the toolbox" for disease scientists. In this context, a disease scientist is an expert with extensive knowledge of a particular disease who seeks ways to prevent or treat the disease. Extensive understanding of a disease enables researchers to look in the "right" place, which may sometimes require employing specific tools, such as the eQTL Catalogue. Typically, pharmaceutical companies aiming to reduce costs and enhance the success rate of potential treatments have teams dedicated to identifying potential targets for a specific disease and other teams to evaluate these targets in various categories (e.g. tractability, toxicity). These companies utilise the eQTL Catalogue either directly or indirectly (through the Open Targets Genetics Platform) for this purpose. I believe this approach is more appropriate than us (as creators of the resource) attempting the same without specific disease expertise. Consequently, the greatest value of this resource lies in its ease of use for researchers other than ourselves.

BIBLIOGRAPHY

[1] F. Crick, Central dogma of molecular biology, Nature. 227 (1970) 561–563.

[2] V. Volloch, Protein-Encoding RNA-to-RNA Information Transfer in Mammalian Cells: Principles of RNA-Dependent mRNA Amplification, Ann Integr Mol Med. 1 (2019). https://doi.org/10.20944/preprints201902.0172.v2.

[3] N.V. Botchkareva, The Molecular Revolution in Cutaneous Biology: Noncoding RNAs: New Molecular Players in Dermatology and Cutaneous Biology, J. Invest. Dermatol. 137 (2017) e105–e111.

[4] K. Kako, J.-D. Kim, A. Fukamizu, Emerging impacts of biological methylation on genetic information, J. Biochem. 165 (2019) 9–18.

[5] T. Schneider-Poetsch, M. Yoshida, Along the Central Dogma-Controlling Gene Expression with Small Molecules, Annu. Rev. Biochem. 87 (2018) 391–420.

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

Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, J. Szustakowki, International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature. 409 (2001) 860–921.

[7] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guigó, M.J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang,

M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, X. Zhu, The sequence of the human genome, Science. 291 (2001) 1304–1351.
[8] A.J. Marian, Sequencing your genome: what does it mean?, Methodist Debakey Cardiovasc. J. 10 (2014) 3–6.
[9] M.W. Hahn, Molecular Population Genetics, Oxford University Press, 2018.
[10] M.V. Suntsova, A.A. Buzdin, Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species, BMC Genomics. 21 (2020) 535.
[11] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature. 526 (2015) 68–74.
[12] T.A. Manolio, Genomewide association studies and assessment of the risk of disease, N. Engl. J. Med. 363 (2010) 166–176.
[13] T.A. Manolio, L.D. Brooks, F.S. Collins, A HapMap harvest of insights into the genetics of common disease, J. Clin. Invest. 118 (2008) 1590–1605.
[14] P.M. Visscher, M.A. Brown, M.I. McCarthy, J. Yang, Five years of GWAS discovery, Am. J. Hum. Genet. 90 (2012) 7–24.
[15] M. Kanai, J.C. Ulirsch, J. Karjalainen, M. Kurki, K.J. Karczewski, E. Fauman, Q.S. Wang, H. Jacobs, F. Aguet, K.G. Ardlie, N. Kerimov, K. Alasoo, C. Benner, K. Ishigaki, S. Sakaue, S. Reilly, Y. Kamatani, K. Matsuda, A. Palotie, B.M. Neale, R. Tewhey, P.C. Sabeti, Y. Okada, M.J. Daly, H.K. Finucane, The BioBank Japan Project, FinnGen, Insights from complex trait fine-mapping across diverse populations, (2021). https://doi.org/10.1101/2021.09.03.21262975.
[16] J.-B. Pingault, F. Rijsdijk, T. Schoeler, S.W. Choi, S. Selzam, E. Krapohl, P.F. O'Reilly, F. Dudbridge, Genetic sensitivity analysis: Adjusting for genetic confounding in epidemiological associations, PLoS Genet. 17 (2021) e1009590.
[17] 05 - Genetic Variation, Google Docs. (n.d.). https://docs.google.com/presentation/d/1JnBiaGG_eJAb1LGUiNaXP4DX-oZKaxaDVg1KFqYeAfA/edit (accessed May 28, 2023).
[18] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M. Stephens, C.D. Bustamante, Genes mirror geography within Europe, Nature. 456 (2008) 98–101.
[19] M. Simcoe, A. Valdes, F. Liu, N.A. Furlotte, D.M. Evans, G. Hemani, S.M. Ring, G.D. Smith, D.L. Duffy, G. Zhu, S.D. Gordon, S.E. Medland, D. Vuckovic, G. Girotto, C. Sala, E. Catamo, M.P. Concas, M. Brumat, P. Gasparini, D. Toniolo, M. Cocca, A. Robino, S. Yazar, A. Hewitt, W. Wu, P. Kraft, C.J. Hammond, Y. Shi, Y. Chen, C. Zeng, C.C.W. Klaver, A.G. Uitterlinden, M.A. Ikram, M.A. Hamer, C.M. van Duijn, T. Nijsten, J. Han, D.A. Mackey, N.G. Martin, C.-Y. Cheng, 23andMe Research Team, International Visible Trait Genetics Consortium, D.A. Hinds, T.D. Spector, M. Kayser, P.G. Hysi, Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color, Sci Adv. 7 (2021). https://doi.org/10.1126/sciadv.abd1239.
[20] L. Yengo, S. Vedantam, E. Marouli, J. Sidorenko, E. Bartell, S. Sakaue, M. Graff, A.U. Eliasen, Y. Jiang, S. Raghavan, J. Miao, J.D. Arias, S.E. Graham, R.E. Mukamel, C.N. Spracklen, X. Yin, S.-H. Chen, T. Ferreira, H.H. Highland, Y. Ji, T. Karaderi, K. Lin, K. Lüll, D.E. Malden, C. Medina-Gomez, M. Machado, A. Moore, S. Rüeger, X. Sim, S. Vrieze, T.S. Ahluwalia, M. Akiyama,

M.A. Allison, M. Alvarez, M.K. Andersen, A. Ani, V. Appadurai, L. Arbeeva, S. Bhaskar, L.F. Bielak, S. Bollepalli, L.L. Bonnycastle, J. Bork-Jensen, J.P. Bradfield, Y. Bradford, P.S. Braund, J.A. Brody, K.S. Burgdorf, B.E. Cade, H. Cai, Q. Cai, A. Campbell, M. Cañadas-Garre, E. Catamo, J.-F. Chai, X. Chai, L.-C. Chang, Y.-C. Chang, C.-H. Chen, A. Chesi, S.H. Choi, R.-H. Chung, M. Cocca, M.P. Concas, C. Couture, G. Cuellar-Partida, R. Danning, E.W. Daw, F. Degenhard, G.E. Delgado, A. Delitala, A. Demirkan, X. Deng, P. Devineni, A. Dietl, M. Dimitriou, L. Dimitrov, R. Dorajoo, A.B. Ekici, J.E. Engmann, Z. Fairhurst-Hunter, A.-E. Farmaki, J.D. Faul, J.-C. Fernandez-Lopez, L. Forer, M. Francescatto, S. Freitag-Wolf, C. Fuchsberger, T.E. Galesloot, Y. Gao, Z. Gao, F. Geller, O. Giannakopoulou, F. Giulianini, A.P. Gjesing, A. Goel, S.D. Gordon, M. Gorski, J. Grove, X. Guo, S. Gustafsson, J. Haessler, T.F. Hansen, A.S. Havulinna, S.J. Haworth, J. He, N. Heard-Costa, P. Hebbar, G. Hindy, Y.-L.A. Ho, E. Hofer, E. Holliday, K. Horn, W.E. Hornsby, J.-J. Hottenga, H. Huang, J. Huang, A. Huerta-Chagoya, J.E. Huffman, Y.-J. Hung, S. Huo, M.Y. Hwang, H. Iha, D.D. Ikeda, M. Isono, A.U. Jackson, S. Jäger, I.E. Jansen, I. Johansson, J.B. Jonas, A. Jonsson, T. Jørgensen, I.-P. Kalafati, M. Kanai, S. Kanoni, L.L. Kårhus, A. Kasturiratne, T. Katsuya, T. Kawaguchi, R.L. Kember, K.A. Kentistou, H.-N. Kim, Y.J. Kim, M.E. Kleber, M.J. Knol, A. Kurbasic, M. Lauzon, P. Le, R. Lea, J.-Y. Lee, H.L. Leonard, S.A. Li, X. Li, X. Li, J. Liang, H. Lin, S.-Y. Lin, J. Liu, X. Liu, K.S. Lo, J. Long, L. Lores-Motta, J. 'an Luan, V. Lyssenko, L.-P. Lyytikäinen, A. Mahajan, V. Mamakou, M. Mangino, A. Manichaikul, J. Marten, M. Mattheisen, L. Mavarani, A.F. McDaid, K. Meidtner, T.L. Melendez, J.M. Mercader, Y. Milaneschi, J.E. Miller, I.Y. Millwood, P.P. Mishra, R.E. Mitchell, L.T. Møllehave, A. Morgan, S. Mucha, M. Munz, M. Nakatochi, C.P. Nelson, M. Nethander, C.W. Nho, A.A. Nielsen, I.M. Nolte, S.S. Nongmaithem, R. Noordam, I. Ntalla, T. Nutile, A. Pandit, P. Christofidou, K. Pärna, M. Pauper, E.R.B. Petersen, L.V. Petersen, N. Pitkänen, O. Polašek, A. Poveda, M.H. Preuss, S. Pyarajan, L.M. Raffield, H. Rakugi, J. Ramirez, A. Rasheed, D. Raven, N.W. Rayner, C. Riveros, R. Rohde, D. Ruggiero, S.E. Ruotsalainen, K.A. Ryan, M. Sabater-Lleal, R. Saxena, M. Scholz, A. Sendamarai, B. Shen, J. Shi, J.H. Shin, C. Sidore, C.M. Sitlani, R.C. Slieker, R.A.J. Smit, A.V. Smith, J.A. Smith, L.J. Smyth, L. Southam, V. Steinthorsdottir, L. Sun, F. Takeuchi, D.S.P. Tallapragada, K.D. Taylor, B.O. Tayo, C. Tcheandjieu, N. Terzikhan, P. Tesolin, A. Teumer, E. Theusch, D.J. Thompson, G. Thorleifsson, P.R.H.J. Timmers, S. Trompet, C. Turman, S. Vaccargiu, S.W. van der Laan, P.J. van der Most, J.B. van Klinken, J. van Setten, S.S. Verma, N. Verweij, Y. Veturi, C.A. Wang, C. Wang, L. Wang, Z. Wang, H.R. Warren, W. Bin Wei, A.R. Wickremasinghe, M. Wielscher, K.L. Wiggins, B.S. Winsvold, A. Wong, Y. Wu, M. Wuttke, R. Xia, T. Xie, K. Yamamoto, J. Yang, J. Yao, H. Young, N.A. Yousri, L. Yu, L. Zeng, W. Zhang, X. Zhang, J.-H. Zhao, W. Zhao, W. Zhou, M.E. Zimmermann, M. Zoledziewska, L.S. Adair, H.H.H. Adams, C.A. Aguilar-Salinas, F. Al-Mulla, D.K. Arnett, F.W. Asselbergs, B.O. Åsvold, J. Attia, B. Banas, S. Bandinelli, D.A. Bennett, T. Bergler, D. Bharadwaj, G. Biino, H. Bisgaard, E. Boerwinkle, C.A. Böger, K. Bønnelykke, D.I. Boomsma, A.D. Børglum, J.B. Borja, C. Bouchard, D.W. Bowden, I. Brandslund, B. Brumpton, J.E. Buring, M.J. Caulfield, J.C. Chambers, G.R. Chandak, S.J. Chanock, N. Chaturvedi, Y.-D.I. Chen, Z. Chen, C.-Y. Cheng, I.E. Christophersen, M. Ciullo, J.W. Cole, F.S. Collins, R.S.

Cooper, M. Cruz, F. Cucca, L.A. Cupples, M.J. Cutler, S.M. Damrauer, T.M. Dantoft, G.J. de Borst, L.C.P.G.M. de Groot, P.L. De Jager, D.P.V. de Kleijn, H. Janaka de Silva, G.V. Dedoussis, A.I. den Hollander, S. Du, D.F. Easton, P.J.M. Elders, A.H. Eliassen, P.T. Ellinor, S. Elmståhl, J. Erdmann, M.K. Evans, D. Fatkin, B. Feenstra, M.F. Feitosa, L. Ferrucci, I. Ford, M. Fornage, A. Franke, P.W. Franks, B.I. Freedman, P. Gasparini, C. Gieger, G. Girotto, M.E. Goddard, Y.M. Golightly, C. Gonzalez-Villalpando, P. Gordon-Larsen, H. Grallert, S.F.A. Grant, N. Grarup, L. Griffiths, V. Gudnason, C. Haiman, H. Hakonarson, T. Hansen, C.A. Hartman, A.T. Hattersley, C. Hayward, S.R. Heckbert, C.-K. Heng, C. Hengstenberg, A.W. Hewitt, H. Hishigaki, C.B. Hoyng, P.L. Huang, W. Huang, S.C. Hunt, K. Hveem, E. Hyppönen, W.G. Iacono, S. Ichihara, M.A. Ikram, C.R. Isasi, R.D. Jackson, M.-R. Jarvelin, Z.-B. Jin, K.-H. Jöckel, P.K. Joshi, P. Jousilahti, J.W. Jukema, M. Kähönen, Y. Kamatani, K.D. Kang, J. Kaprio, S.L.R. Kardia, F. Karpe, N. Kato, F. Kee, T. Kessler, A.V. Khera, C.C. Khor, L.A.L.M. Kiemeney, B.-J. Kim, E.K. Kim, H.-L. Kim, P. Kirchhof, M. Kivimaki, W.-P. Koh, H.A. Koistinen, G.D. Kolovou, J.S. Kooner, C. Kooperberg, A. Köttgen, P. Kovacs, A. Kraaijeveld, P. Kraft, R.M. Krauss, M. Kumari, Z. Kutalik, M. Laakso, L.A. Lange, C. Langenberg, L.J. Launer, L. Le Marchand, H. Lee, N.R. Lee, T. Lehtimäki, H. Li, L. Li, W. Lieb, X. Lin, L. Lind, A. Linneberg, C.-T. Liu, J. Liu, M. Loeffler, B. London, S.A. Lubitz, S.J. Lye, D.A. Mackey, R. Mägi, P.K.E. Magnusson, G.M. Marcus, P.M. Vidal, N.G. Martin, W. März, F. Matsuda, R.W. McGarrah, M. McGue, A.J. McKnight, S.E. Medland, D. Mellström, A. Metspalu, B.D. Mitchell, P. Mitchell, D.O. Mook-Kanamori, A.D. Morris, L.A. Mucci, P.B. Munroe, M.A. Nalls, S. Nazarian, A.E. Nelson, M.J. Neville, C. Newton-Cheh, C.S. Nielsen, M.M. Nöthen, C. Ohlsson, A.J. Oldehinkel, L. Orozco, K. Pahkala, P. Pajukanta, C.N.A. Palmer, E.J. Parra, C. Pattaro, O. Pedersen, C.E. Pennell, B.W.J.H. Penninx, L. Perusse, A. Peters, P.A. Peyser, D.J. Porteous, D. Posthuma, C. Power, P.P. Pramstaller, M.A. Province, Q. Qi, J. Qu, D.J. Rader, O.T. Raitakari, S. Ralhan, L.S. Rallidis, D.C. Rao, S. Redline, D.F. Reilly, A.P. Reiner, S.Y. Rhee, P.M. Ridker, M. Rienstra, S. Ripatti, M.D. Ritchie, D.M. Roden, F.R. Rosendaal, J.I. Rotter, I. Rudan, F. Rutters, C. Sabanayagam, D. Saleheen, V. Salomaa, N.J. Samani, D.K. Sanghera, N. Sattar, B. Schmidt, H. Schmidt, R. Schmidt, M.B. Schulze, H. Schunkert, L.J. Scott, R.J. Scott, P. Sever, E.J. Shiroma, M.B. Shoemaker, X.-O. Shu, E.M. Simonsick, M. Sims, J.R. Singh, A.B. Singleton, M.F. Sinner, J.G. Smith, H. Snieder, T.D. Spector, M.J. Stampfer, K.J. Stark, D.P. Strachan, L.M. ’t Hart, Y. Tabara, H. Tang, J.-C. Tardif, T.A. Thanaraj, N.J. Timpson, A. Tönjes, A. Tremblay, T. Tuomi, J. Tuomilehto, M.-T. Tusié-Luna, A.G. Uitterlinden, R.M. van Dam, P. van der Harst, N. Van der Velde, C.M. van Duijn, N.M. van Schoor, V. Vitart, U. Völker, P. Vollenweider, H. Völzke, N.H. Wacher-Rodarte, M. Walker, Y.X. Wang, N.J. Wareham, R.M. Watanabe, H. Watkins, D.R. Weir, T.M. Werge, E. Widen, L.R. Wilkens, G. Willemsen, W.C. Willett, J.F. Wilson, T.-Y. Wong, J.-T. Woo, A.F. Wright, J.-Y. Wu, H. Xu, C.S. Yajnik, M. Yokota, J.-M. Yuan, E. Zeggini, B.S. Zemel, W. Zheng, X. Zhu, J.M. Zmuda, A.B. Zonderman, J.-A. Zwart, 23andMe Research Team, VA Million Veteran Program, DiscovEHR (DiscovEHR and MyCode Community Health Initiative), eMERGE (Electronic Medical Records and Genomics Network), Lifelines Cohort Study, PRACTICAL Consortium, Understanding Society Scientific Group, D.I. Chasman, Y.S. Cho, I.M. Heid, M.I. McCarthy, M.C.Y.

Ng, C.J. O'Donnell, F. Rivadeneira, U. Thorsteinsdottir, Y.V. Sun, E.S. Tai, M. Boehnke, P. Deloukas, A.E. Justice, C.M. Lindgren, R.J.F. Loos, K.L. Mohlke, K.E. North, K. Stefansson, R.G. Walters, T.W. Winkler, K.L. Young, P.-R. Loh, J. Yang, T. Esko, T.L. Assimes, A. Auton, G.R. Abecasis, C.J. Willer, A.E. Locke, S.I. Berndt, G. Lettre, T.M. Frayling, Y. Okada, A.R. Wood, P.M. Visscher, J.N. Hirschhorn, A saturated map of common genetic variants associated with human height, Nature. 610 (2022) 704–712.

[21] V.L. Chen, X. Du, Y. Chen, A. Kuppa, S.K. Handelman, R.B. Vohnoutka, P.A. Peyser, N.D. Palmer, L.F. Bielak, B. Halligan, E.K. Speliotes, Genome-wide association study of serum liver enzymes implicates diverse metabolic and liver pathology, Nat. Commun. 12 (2021) 816.

[22] K.S. Ruth, P.J. Campbell, S. Chew, E.M. Lim, N. Hadlow, B.G.A. Stuckey, S.J. Brown, B. Feenstra, J. Joseph, G.L. Surdulescu, H.F. Zheng, J.B. Richards, A. Murray, T.D. Spector, S.G. Wilson, J.R.B. Perry, Genome-wide association study with 1000 genomes imputation identifies signals for nine sex hormone-related phenotypes, Eur. J. Hum. Genet. 24 (2016) 284–290.

[23] Y. Liu, A code within the genetic code: codon usage regulates co-translational protein folding, Cell Commun. Signal. 18 (2020) 145.

[24] M.K. Sakharkar, V.T.K. Chow, I. Chaturvedi, V.S. Mathura, P. Shapshak, P. Kangueane, A report on single exon genes (SEG) in eukaryotes, Front. Biosci. 9 (2004) 3262–3267.

[25] Y. Lee, D.C. Rio, Mechanisms and Regulation of Alternative Pre-mRNA Splicing, Annu. Rev. Biochem. 84 (2015) 291–323.

[26] Y. Barash, J.A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B.J. Blencowe, B.J. Frey, Deciphering the splicing code, Nature. 465 (2010) 53–59.

[27] A. Kalsotra, T.A. Cooper, Functional consequences of developmentally regulated alternative splicing, Nat. Rev. Genet. 12 (2011) 715–729.

[28] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, C.B. Burge, Alternative isoform regulation in human tissue transcriptomes, Nature. 456 (2008) 470–476.

[29] J.L. Aspden, E.W.J. Wallace, N. Whiffin, Not all exons are protein coding: Addressing a common misconception, Cell Genom. 3 (2023) 100296.

[30] X. Jia, Y. Yang, Y. Chen, Z. Xia, W. Zhang, Y. Feng, Y. Li, J. Tan, C. Xu, Q. Zhang, H. Deng, X. Shi, Multivariate analysis of genome-wide data to identify potential pleiotropic genes for type 2 diabetes, obesity and coronary artery disease using MetaCCA, Int. J. Cardiol. 283 (2019) 144–150.

[31] J.P. Bradfield, H.R. Taal, N.J. Timpson, A. Scherag, C. Lecoeur, N.M. Warrington, E. Hypponen, C. Holst, B. Valcarcel, E. Thiering, R.M. Salem, F.R. Schumacher, D.L. Cousminer, P.M.A. Sleiman, J. Zhao, R.I. Berkowitz, K.S. Vimaleswaran, I. Jarick, C.E. Pennell, D.M. Evans, B. St Pourcain, D.J. Berry, D.O. Mook-Kanamori, A. Hofman, F. Rivadeneira, A.G. Uitterlinden, C.M. van Duijn, R.J.P. van der Valk, J.C. de Jongste, D.S. Postma, D.I. Boomsma, W.J. Gauderman, M.T. Hassanein, C.M. Lindgren, R. Mägi, C.A.G. Boreham, C.E. Neville, L.A. Moreno, P. Elliott, A. Pouta, A.-L. Hartikainen, M. Li, O. Raitakari, T. Lehtimäki, J.G. Eriksson, A. Palotie, J. Dallongeville, S. Das, P. Deloukas, G. McMahon, S.M. Ring, J.P. Kemp, J.L. Buxton, A.I.F. Blakemore, M. Bustamante, M. Guxens, J.N. Hirschhorn, M.W. Gillman, E. Kreiner-Møller, H. Bisgaard, F.D. Gilliland, J. Heinrich, E. Wheeler, I. Barroso, S. O'Rahilly, A. Meirhaeghe, T.I.A. Sørensen, C. Power, L.J. Palmer, A. Hinney, E. Widen, I.S.

Farooqi, M.I. McCarthy, P. Froguel, D. Meyre, J. Hebebrand, M.-R. Jarvelin, V.W.V. Jaddoe, G.D. Smith, H. Hakonarson, S.F.A. Grant, Early Growth Genetics Consortium, A genome-wide association meta-analysis identifies new childhood obesity loci, Nat. Genet. 44 (2012) 526–531.

[32] G. Donati, I. Dumontheil, E.L. Meaburn, Genome-Wide Association Study of Latent Cognitive Measures in Adolescence: Genetic Overlap With Intelligence and Education, Mind Brain Educ. 13 (2019) 224–233.

[33] D. Zabaneh, E. Krapohl, H.A. Gaspar, C. Curtis, S.H. Lee, H. Patel, S. Newhouse, H.M. Wu, M.A. Simpson, M. Putallaz, D. Lubinski, R. Plomin, G. Breen, A genome-wide association study for extremely high intelligence, Mol. Psychiatry. 23 (2018) 1226–1232.

[34] X. Yin, L.S. Chan, D. Bose, A.U. Jackson, P. VandeHaar, A.E. Locke, C. Fuchsberger, H.M. Stringham, R. Welch, K. Yu, L. Fernandes Silva, S.K. Service, D. Zhang, E.C. Hector, E. Young, L. Ganel, I. Das, H. Abel, M.R. Erdos, L.L. Bonnycastle, J. Kuusisto, N.O. Stitziel, I.M. Hall, G.R. Wagner, FinnGen, J. Kang, J. Morrison, C.F. Burant, F.S. Collins, S. Ripatti, A. Palotie, N.B. Freimer, K.L. Mohlke, L.J. Scott, X. Wen, E.B. Fauman, M. Laakso, M. Boehnke, Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci, Nat. Commun. 13 (2022) 1644.

[35] P.G. Hysi, M. Mangino, P. Christofidou, M. Falchi, E.D. Karoly, Nihr Bioresource Investigators, R.P. Mohney, A.M. Valdes, T.D. Spector, C. Menni, Metabolome Genome-Wide Association Study Identifies 74 Novel Genomic Regions Influencing Plasma Metabolites Levels, Metabolites. 12 (2022). https://doi.org/10.3390/metabo12010061.

[36] F. Aguet, K. Alasoo, Y.I. Li, A. Battle, H.K. Im, S.B. Montgomery, T. Lappalainen, Molecular quantitative trait loci, Nature Reviews Methods Primers. 3 (2023) 1–22.

[37] Y. Tanigawa, J. Qian, G. Venkataraman, J.M. Justesen, R. Li, R. Tibshirani, T. Hastie, M.A. Rivas, Significant sparse polygenic risk scores across 813 traits in UK Biobank, PLoS Genet. 18 (2022) e1010105.

[38] G. Breen, Q. Li, B.L. Roth, P. O'Donnell, M. Didriksen, R. Dolmetsch, P.F. O'Reilly, H.A. Gaspar, H. Manji, C. Huebel, J.R. Kelsoe, D. Malhotra, A. Bertolino, D. Posthuma, P. Sklar, S. Kapur, P.F. Sullivan, D.A. Collier, H.J. Edenberg, Translating genome-wide association findings into new therapeutics for psychiatry, Nat. Neurosci. 19 (2016) 1392–1396.

[39] M.T. Maurano, R. Humbert, E. Rynes, R.E. Thurman, E. Haugen, H. Wang, A.P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A.K. Johnson, T.K. Canfield, E. Giste, M. Diegel, D. Bates, R.S. Hansen, S. Neph, P.J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S.R. Sunyaev, R. Kaul, J.A. Stamatoyannopoulos, Systematic localization of common disease-associated variation in regulatory DNA, Science. 337 (2012) 1190–1195.

[40] J.A. McQuerry, M. Mclaird, S.N. Hartin, J.C. Means, J. Johnston, T. Pastinen, S.T. Younger, Massively parallel identification of functionally consequential noncoding genetic variants in undiagnosed rare disease patients, Sci. Rep. 12 (2022) 7576.

[41] K. Watanabe, E. Taskesen, A. van Bochoven, D. Posthuma, Functional mapping and annotation of genetic associations with FUMA, Nat. Commun. 8 (2017) 1826.

[42] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 9362–9367.
[43] D.L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M.E. Dolan, N.J. Cox, Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS, PLoS Genet. 6 (2010) e1000888.
[44] N. Weinhold, A. Jacobsen, N. Schultz, C. Sander, W. Lee, Genome-wide analysis of noncoding regulatory mutations in cancer, Nat. Genet. 46 (2014) 1160–1165.
[45] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature. 489 (2012) 57–74.
[46] R.C. Jansen, J.P. Nap, Genetical genomics: the added value from segregation, Trends Genet. 17 (2001) 388–391.
[47] R.B. Brem, G. Yvert, R. Clinton, L. Kruglyak, Genetic dissection of transcriptional regulation in budding yeast, Science. 296 (2002) 752–755.
[48] U. Võsa, A. Claringbould, H.-J. Westra, M.J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Yazar, H. Brugge, R. Oelen, D.H. de Vries, M.G.P. van der Wijst, S. Kasela, N. Pervjakova, I. Alves, M.-J. Favé, M. Agbessi, M.W. Christiansen, R. Jansen, I. Seppälä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. Sönmez Flitman, A. Brown, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg, J. Kettunen, B. Lee, F. Zhang, T. Qi, J.A. Hernandez, W. Arindrarto, F. Beutner, J. Dmitrieva, M. Elansary, B.P. Fairfax, M. Georges, B.T. Heijmans, A.W. Hewitt, M. Kähönen, Y. Kim, J.C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U.M. Marigorta, H. Mei, Y. Momozawa, M. Müller-Nurasyid, M. Nauck, M.G. Nivard, B.W.J.H. Penninx, J.K. Pritchard, O.T. Raitakari, O. Rotzschke, E.P. Slagboom, C.D.A. Stehouwer, M. Stumvoll, P. Sullivan, P.A.C. 't Hoen, J. Thiery, A. Tönjes, J. van Dongen, M. van Iterson, J.H. Veldink, U. Völker, R. Warmerdam, C. Wijmenga, M. Swertz, A. Andiappan, G.W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, E. Dermitzakis, S. Bergmann, T. Frayling, J. van Meurs, H. Prokisch, H. Ahsan, B.L. Pierce, T. Lehtimäki, D.I. Boomsma, B.M. Psaty, S.A. Gharib, P. Awadalla, L. Milani, W.H. Ouwehand, K. Downes, O. Stegle, A. Battle, P.M. Visscher, J. Yang, M. Scholz, J. Powell, G. Gibson, T. Esko, L. Franke, Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression, Nat. Genet. (2021) 1–11.
[49] E.B. Fauman, C. Hyde, An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs, BMC Bioinformatics. 23 (2022) 169.
[50] Y. Gilad, S.A. Rifkin, J.K. Pritchard, Revealing the architecture of gene regulation: the promise of eQTL studies, Trends Genet. 24 (2008) 408–415.
[51] GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study,

Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, A. Battle, C.D. Brown, B.E. Engelhardt, S.B. Montgomery, Genetic effects on gene expression across human tissues, Nature. 550 (2017) 204–213.

[52] H.-J. Westra, L. Franke, From genome to function by studying eQTLs, Biochim. Biophys. Acta. 1842 (2014) 1896–1902.

[53] L. Kolberg, Developing and applying bioinformatics tools for gene expression data interpretation[No title], PhD, University of Tartu, 2021. https://dspace.ut.ee/handle/10062/71972 (accessed May 30, 2023).

[54] D.W. Yao, L.J. O'Connor, A.L. Price, A. Gusev, Quantifying genetic effects on disease mediated by assayed gene expression levels, Nat. Genet. 52 (2020) 626–633.

[55] T. Qi, Y. Wu, H. Fang, F. Zhang, S. Liu, J. Zeng, J. Yang, Genetic control of RNA splicing and its distinct role in complex trait variation, Nat. Genet. 54 (2022) 1355–1363.

[56] B. Zeng, J. Bendl, R. Kosoy, J.F. Fullard, G.E. Hoffman, P. Roussos, Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits, Nat. Genet. 54 (2022) 161–169.

[57] H. Guo, M.D. Fortune, O.S. Burren, E. Schofield, J.A. Todd, C. Wallace, Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases, Hum. Mol. Genet. 24 (2015) 3305–3313.

[58] M. Wainberg, N. Sinnott-Armstrong, N. Mancuso, A.N. Barbeira, D.A. Knowles, D. Golan, R. Ermel, A. Ruusalepp, T. Quertermous, K. Hao, J.L.M. Björkegren, H.K. Im, B. Pasaniuc, M.A. Rivas, A. Kundaje, Opportunities and challenges for transcriptome-wide association studies, Nat. Genet. 51 (2019) 592–599.

[59] M.E. Hauberg, W. Zhang, C. Giambartolomei, O. Franzén, D.L. Morris, T.J. Vyse, A. Ruusalepp, CommonMind Consortium, P. Sklar, E.E. Schadt, J.L.M. Björkegren, P. Roussos, Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression, Am. J. Hum. Genet. 101 (2017) 157.

[60] X. Wen, R. Pique-Regi, F. Luca, Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization, PLoS Genet. 13 (2017) e1006646.

[61] F. Hormozdiari, M. van de Bunt, A.V. Segrè, X. Li, J.W.J. Joo, M. Bilow, J.H. Sul, S. Sankararaman, B. Pasaniuc, E. Eskin, Colocalization of GWAS and eQTL Signals Detects Target Genes, Am. J. Hum. Genet. 99 (2016) 1245–1260.

[62] C. Giambartolomei, D. Vukcevic, E.E. Schadt, L. Franke, A.D. Hingorani, C. Wallace, V. Plagnol, Bayesian test for colocalisation between pairs of genetic association studies using summary statistics, PLoS Genet. 10 (2014) e1004383.

[63] C. Wallace, Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses, PLoS Genet. 16 (2020) e1008720.

[64] C. Giambartolomei, J. Zhenli Liu, W. Zhang, M. Hauberg, H. Shi, J. Boocock, J. Pickrell, A.E. Jaffe, CommonMind Consortium, B. Pasaniuc, P. Roussos, A Bayesian framework for multiple trait colocalization from summary association statistics, Bioinformatics. 34 (2018) 2538–2545.

[65] Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M.R. Robinson, J.E. Powell, G.W. Montgomery, M.E. Goddard, N.R. Wray, P.M. Visscher, J. Yang, Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets, Nat. Genet. 48 (2016) 481–487.
[66] G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping, J. R. Stat. Soc. Series B Stat. Methodol. 82 (2020) 1273–1300.
[67] C. Wallace, A more accurate method for colocalisation analysis allowing for multiple causal variants, PLoS Genet. 17 (2021) e1009440.
[68] A. Buil, A.A. Brown, T. Lappalainen, A. Viñuela, M.N. Davies, H.-F. Zheng, J.B. Richards, D. Glass, K.S. Small, R. Durbin, T.D. Spector, E.T. Dermitzakis, Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins, Nat. Genet. 47 (2015) 88–91.
[69] Q.S. Wang, H. Huang, Methods for statistical fine-mapping and their applications to auto-immune diseases, Semin. Immunopathol. 44 (2022) 101–113.
[70] E. Mountjoy, E.M. Schmidt, M. Carmona, J. Schwartzentruber, G. Peat, A. Miranda, L. Fumis, J. Hayhurst, A. Buniello, M.A. Karim, D. Wright, A. Hercules, E. Papa, E.B. Fauman, J.C. Barrett, J.A. Todd, D. Ochoa, I. Dunham, M. Ghoussaini, An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci, Nat. Genet. 53 (2021) 1527–1533.
[71] W. Chen, B.R. Larrabee, I.G. Ovsyannikova, R.B. Kennedy, I.H. Haralambieva, G.A. Poland, D.J. Schaid, Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics, Genetics. 200 (2015) 719–736.
[72] F. Hormozdiari, E. Kostem, E.Y. Kang, B. Pasaniuc, E. Eskin, Identifying causal variants at loci with multiple signals of association, Genetics. 198 (2014) 497–508.
[73] C. Benner, C.C.A. Spencer, A.S. Havulinna, V. Salomaa, S. Ripatti, M. Pirinen, FINEMAP: efficient variable selection using summary data from genome-wide association studies, Bioinformatics. 32 (2016) 1493–1501.
[74] P.J. Newcombe, D.V. Conti, S. Richardson, JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects, Genet. Epidemiol. 40 (2016) 188–201.
[75] Y. Zou, P. Carbonetto, G. Wang, M. Stephens, Fine-mapping from summary data with the “Sum of Single Effects” model, PLoS Genet. 18 (2022) e1010299.
[76] A.S. Dimas, S. Deutsch, B.E. Stranger, S.B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M.G. Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E.T. Dermitzakis, S.E. Antonarakis, Common Regulatory Variation Impacts Gene Expression in a Cell Type–Dependent Manner, Science. 325 (2009) 1246–1250.
[77] X. Liu, H.K. Finucane, A. Gusev, G. Bhatia, S. Gazal, L. O’Connor, B. Bulik-Sullivan, F.A. Wright, P.F. Sullivan, B.M. Neale, A.L. Price, Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues, Am. J. Hum. Genet. 100 (2017) 605–616.
[78] Z. Mu, W. Wei, B. Fair, J. Miao, P. Zhu, Y.I. Li, The impact of cell type and context-dependent regulatory variants on human immune traits, Genome Biol. 22 (2021) 122.

[79] K. Alasoo, J. Rodrigues, J. Danesh, D.F. Freitag, D.S. Paul, D.J. Gaffney, Genetic effects on promoter usage are highly context-specific and contribute to complex traits, Elife. 8 (2019). https://doi.org/10.7554/eLife.41673.
[80] L.E. McNamara, R.J. McMurray, M.J.P. Biggs, F. Kantawong, R.O.C. Oreffo, M.J. Dalby, Nanotopographical control of stem cell differentiation, J. Tissue Eng. 2010 (2010) 120623.
[81] J.M.W. Slack, Metaplasia and transdifferentiation: from pure biology to the clinic, Nat. Rev. Mol. Cell Biol. 8 (2007) 369–378.
[82] N. Yosef, A. Regev, Writ large: Genomic dissection of the effect of cellular environment on immune response, Science. 354 (2016) 64–68.
[83] M.D. Gallagher, A.S. Chen-Plotkin, The Post-GWAS Era: From Association to Function, Am. J. Hum. Genet. 102 (2018) 717–730.
[84] N. Aizarani, A. Saviano, Sagar, L. Mailly, S. Durand, J.S. Herman, P. Pessaux, T.F. Baumert, D. Grün, A human liver cell atlas reveals heterogeneity and epithelial progenitors, Nature. 572 (2019) 199–204.
[85] Tabula Sapiens Consortium*, R.C. Jones, J. Karkanias, M.A. Krasnow, A.O. Pisco, S.R. Quake, J. Salzman, N. Yosef, B. Bulthaup, P. Brown, W. Harper, M. Hemenez, R. Ponnusamy, A. Salehi, B.A. Sanagavarapu, E. Spallino, K.A. Aaron, W. Concepcion, J.M. Gardner, B. Kelly, N. Neidlinger, Z. Wang, S. Crasta, S. Kolluru, M. Morri, A.O. Pisco, S.Y. Tan, K.J. Travaglini, C. Xu, M. Alcántara-Hernández, N. Almanzar, J. Antony, B. Beyersdorf, D. Burhan, K. Calcuttawala, M.M. Carter, C.K.F. Chan, C.A. Chang, S. Chang, A. Colville, S. Crasta, R.N. Culver, I. Cvijović, G. D'Amato, C. Ezran, F.X. Galdos, A. Gillich, W.R. Goodyer, Y. Hang, A. Hayashi, S. Houshdaran, X. Huang, J.C. Irwin, S. Jang, J.V. Juanico, A.M. Kershner, S. Kim, B. Kiss, S. Kolluru, W. Kong, M.E. Kumar, A.H. Kuo, R. Leylek, B. Li, G.B. Loeb, W.-J. Lu, S. Mantri, M. Markovic, P.L. McAlpine, A. de Morree, M. Morri, K. Mrouj, S. Mukherjee, T. Muser, P. Neuhöfer, T.D. Nguyen, K. Perez, R. Phansalkar, A.O. Pisco, N. Puluca, Z. Qi, P. Rao, H. Raquer-McKay, N. Schaum, B. Scott, B. Seddighzadeh, J. Segal, S. Sen, S. Sikandar, S.P. Spencer, L.C. Steffes, V.R. Subramaniam, A. Swarup, M. Swift, K.J. Travaglini, W. Van Treuren, E. Trimm, S. Veizades, S. Vijayakumar, K.C. Vo, S.K. Vorperian, W. Wang, H.N.W. Weinstein, J. Winkler, T.T.H. Wu, J. Xie, A.R. Yung, Y. Zhang, A.M. Detweiler, H. Mekonen, N.F. Neff, R.V. Sit, M. Tan, J. Yan, G.R. Bean, V. Charu, E. Forgó, B.A. Martin, M.G. Ozawa, O. Silva, S.Y. Tan, A. Toland, V.N.P. Vemuri, S. Afik, K. Awayan, O.B. Botvinnik, A. Byrne, M. Chen, R. Dehghannasiri, A.M. Detweiler, A. Gayoso, A.A. Granados, Q. Li, G. Mahmoudabadi, A. McGeever, A. de Morree, J.E. Olivieri, M. Park, A.O. Pisco, N. Ravikumar, J. Salzman, G. Stanley, M. Swift, M. Tan, W. Tan, A.J. Tarashansky, R. Vanheusden, S.K. Vorperian, P. Wang, S. Wang, G. Xing, C. Xu, N. Yosef, M. Alcántara-Hernández, J. Antony, C.K.F. Chan, C.A. Chang, A. Colville, S. Crasta, R. Culver, L. Dethlefsen, C. Ezran, A. Gillich, Y. Hang, P.-Y. Ho, J.C. Irwin, S. Jang, A.M. Kershner, W. Kong, M.E. Kumar, A.H. Kuo, R. Leylek, S. Liu, G.B. Loeb, W.-J. Lu, J.S. Maltzman, R.J. Metzger, A. de Morree, P. Neuhöfer, K. Perez, R. Phansalkar, Z. Qi, P. Rao, H. Raquer-McKay, K. Sasagawa, B. Scott, R. Sinha, H. Song, S.P. Spencer, A. Swarup, M. Swift, K.J. Travaglini, E. Trimm, S. Veizades, S. Vijayakumar, B. Wang, W. Wang, J. Winkler, J. Xie, A.R. Yung, S.E. Artandi, P.A. Beachy, M.F. Clarke, L.C. Giudice, F.W. Huang, K.C. Huang, J. Idoyaga, S.K. Kim, M. Krasnow, C.S.

Kuo, P. Nguyen, S.R. Quake, T.A. Rando, K. Red-Horse, J. Reiter, D.A. Relman, J.L. Sonnenburg, B. Wang, A. Wu, S.M. Wu, T. Wyss-Coray, The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans, Science. 376 (2022) eabl4896.

[86] O. Rozenblatt-Rosen, M.J.T. Stubbington, A. Regev, S.A. Teichmann, The Human Cell Atlas: from vision to reality, Nature Publishing Group UK. (2017). https://doi.org/10.1038/550451a.

[87] P. Xu, M. Wang, W.-M. Song, Q. Wang, G.-C. Yuan, P.H. Sudmant, H. Zare, Z. Tu, M.E. Orr, B. Zhang, The landscape of human tissue and cell type specific expression and co-regulation of senescence genes, Mol. Neurodegener. 17 (2022) 5.

[88] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, Garland Science, 2002.

[89] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, An Overview of Gene Control, Garland Science, 2002.

[90] I. Seim, S. Ma, V.N. Gladyshev, Gene expression signatures of human cell and tissue longevity, NPJ Aging Mech Dis. 2 (2016) 16014.

[91] M. Fagny, J.N. Paulson, M.L. Kuijjer, A.R. Sonawane, C.-Y. Chen, C.M. Lopes-Ramos, K. Glass, J. Quackenbush, J. Platig, Exploring regulation in tissues with eQTL networks, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) E7841–E7850.

[92] K.K.-H. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W.J. Housley, S. Beik, N. Shoresh, H. Whitton, R.J.H. Ryan, A.A. Shishkin, M. Hatan, M.J. Carrasco-Alfonso, D. Mayer, C.J. Luckey, N.A. Patsopoulos, P.L. De Jager, V.K. Kuchroo, C.B. Epstein, M.J. Daly, D.A. Hafler, B.E. Bernstein, Genetic and epigenetic fine mapping of causal autoimmune disease variants, Nature. 518 (2015) 337–343.

[93] F.W. Albert, L. Kruglyak, The role of regulatory variation in complex traits and disease, Nat. Rev. Genet. 16 (2015) 197–212.

[94] A.C. Nica, E.T. Dermitzakis, Using gene expression to investigate the genetic basis of complex disorders, Hum. Mol. Genet. 17 (2008) R129–34.

[95] A.T. McKenzie, M. Wang, M.E. Hauberg, J.F. Fullard, A. Kozlenkov, A. Keenan, Y.L. Hurd, S. Dracheva, P. Casaccia, P. Roussos, B. Zhang, Brain Cell Type Specific Gene Expression and Co-expression Network Architectures, Sci. Rep. 8 (2018) 8868.

[96] N. Kerimov, J.D. Hayhurst, K. Peikova, J.R. Manning, P. Walter, L. Kolberg, M. Samoviča, M.P. Sakthivel, I. Kuzmin, S.J. Trevanion, T. Burdett, S. Jupp, H. Parkinson, I. Papatheodorou, A.D. Yates, D.R. Zerbino, K. Alasoo, A compendium of uniformly processed human gene expression and splicing quantitative trait loci, Nat. Genet. 53 (2021) 1290–1299.

[97] H. Mostafavi, J.P. Spence, S. Naqvi, J.K. Pritchard, Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery, bioRxiv. (2022) 2022.05.07.491045. https://doi.org/10.1101/2022.05.07.491045.

[98] K. Lepik, Inferring causality between transcriptome and complex traits, PhD, University of Tartu, 2021. http://dspace.ut.ee/handle/10062/71645?show=full (accessed April 16, 2023).

[99] G.D. Smith, S. Ebrahim, Mendelian randomization: prospects, potentials, and limitations, Int. J. Epidemiol. 33 (2004) 30–42.

[100] G. Davey Smith, G. Hemani, Mendelian randomization: genetic anchors for causal inference in epidemiological studies, Hum. Mol. Genet. 23 (2014) R89–98.

[101] E.T. Lim, P. Würtz, A.S. Havulinna, P. Palta, T. Tukiainen, K. Rehnström, T. Esko, R. Mägi, M. Inouye, T. Lappalainen, Y. Chan, R.M. Salem, M. Lek, J. Flannick, X. Sim, A. Manning, C. Ladenvall, S. Bumpstead, E. Hämäläinen, K. Aalto, M. Maksimow, M. Salmi, S. Blankenberg, D. Ardissino, S. Shah, B. Horne, R. McPherson, G.K. Hovingh, M.P. Reilly, H. Watkins, A. Goel, M. Farrall, D. Girelli, A.P. Reiner, N.O. Stitziel, S. Kathiresan, S. Gabriel, J.C. Barrett, T. Lehtimäki, M. Laakso, L. Groop, J. Kaprio, M. Perola, M.I. McCarthy, M. Boehnke, D.M. Altshuler, C.M. Lindgren, J.N. Hirschhorn, A. Metspalu, N.B. Freimer, T. Zeller, S. Jalkanen, S. Koskinen, O. Raitakari, R. Durbin, D.G. MacArthur, V. Salomaa, S. Ripatti, M.J. Daly, A. Palotie, Sequencing Initiative Suomi (SISu) Project, Distribution and medical impact of loss-of-function variants in the Finnish founder population, PLoS Genet. 10 (2014) e1004494.

[102] B.F. Voight, G.M. Peloso, M. Orho-Melander, R. Frikke-Schmidt, M. Barbalic, M.K. Jensen, G. Hindy, H. Hólm, E.L. Ding, T. Johnson, H. Schunkert, N.J. Samani, R. Clarke, J.C. Hopewell, J.F. Thompson, M. Li, G. Thorleifsson, C. Newton-Cheh, K. Musunuru, J.P. Pirruccello, D. Saleheen, L. Chen, A.F.R. Stewart, A. Schillert, U. Thorsteinsdottir, G. Thorgeirsson, S. Anand, J.C. Engert, T. Morgan, J. Spertus, M. Stoll, K. Berger, N. Martinelli, D. Girelli, P.P. McKeown, C.C. Patterson, S.E. Epstein, J. Devaney, M.-S. Burnett, V. Mooser, S. Ripatti, I. Surakka, M.S. Nieminen, J. Sinisalo, M.-L. Lokki, M. Perola, A. Havulinna, U. de Faire, B. Gigante, E. Ingelsson, T. Zeller, P. Wild, P.I.W. de Bakker, O.H. Klungel, A.-H. Maitland-van der Zee, B.J.M. Peters, A. de Boer, D.E. Grobbee, P.W. Kamphuisen, V.H.M. Deneer, C.C. Elbers, N.C. Onland-Moret, M.H. Hofker, C. Wijmenga, W.M.M. Verschuren, J.M.A. Boer, Y.T. van der Schouw, A. Rasheed, P. Frossard, S. Demissie, C. Willer, R. Do, J.M. Ordovas, G.R. Abecasis, M. Boehnke, K.L. Mohlke, M.J. Daly, C. Guiducci, N.P. Burtt, A. Surti, E. Gonzalez, S. Purcell, S. Gabriel, J. Marrugat, J. Peden, J. Erdmann, P. Diemert, C. Willenborg, I.R. König, M. Fischer, C. Hengstenberg, A. Ziegler, I. Buysschaert, D. Lambrechts, F. Van de Werf, K.A. Fox, N.E. El Mokhtari, D. Rubin, J. Schrezenmeir, S. Schreiber, A. Schäfer, J. Danesh, S. Blankenberg, R. Roberts, R. McPherson, H. Watkins, A.S. Hall, K. Overvad, E. Rimm, E. Boerwinkle, A. Tybjaerg-Hansen, L.A. Cupples, M.P. Reilly, O. Melander, P.M. Mannucci, D. Ardissino, D. Siscovick, R. Elosua, K. Stefansson, C.J. O'Donnell, V. Salomaa, D.J. Rader, L. Peltonen, S.M. Schwartz, D. Altshuler, S. Kathiresan, Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study, Lancet. 380 (2012) 572–580.

[103] N.M.G. De Silva, R.M. Freathy, T.M. Palmer, L.A. Donnelly, J. 'an Luan, T. Gaunt, C. Langenberg, M.N. Weedon, B. Shields, B.A. Knight, K.J. Ward, M.S. Sandhu, R.M. Harbord, M.I. McCarthy, G.D. Smith, S. Ebrahim, A.T. Hattersley, N. Wareham, D.A. Lawlor, A.D. Morris, C.N.A. Palmer, T.M. Frayling, Mendelian randomization studies do not support a role for raised circulating triglyceride levels influencing type 2 diabetes, glucose levels, or insulin resistance, Diabetes. 60 (2011) 1008–1018.

[104] S. Ebrahim, G. Davey Smith, Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology?, Hum. Genet. 123 (2008) 15–33.
[105] D.A. Lawlor, R.M. Harbord, J.A.C. Sterne, N. Timpson, G. Davey Smith, Mendelian randomization: using genes as instruments for making causal inferences in epidemiology, Stat. Med. 27 (2008) 1133–1163.
[106] E. Porcu, M.C. Sadler, K. Lepik, C. Auwerx, A.R. Wood, A. Weihs, M.S.B. Sleiman, D.M. Ribeiro, S. Bandinelli, T. Tanaka, M. Nauck, U. Völker, O. Delaneau, A. Metspalu, A. Teumer, T. Frayling, F.A. Santoni, A. Reymond, Z. Kutalik, Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome, Nat. Commun. 12 (2021) 5647.
[107] G. McVicker, B. van de Geijn, J.F. Degner, C.E. Cain, N.E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, J.K. Pritchard, Identification of genetic variants that affect histone modifications in human cells, Science. 342 (2013) 747–749.
[108] L. Chen, B. Ge, F.P. Casale, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yan, K. Kundu, S. Ecker, A. Datta, D. Richardson, F. Burden, D. Mead, A.L. Mann, J.M. Fernandez, S. Rowlston, S.P. Wilder, S. Farrow, X. Shao, J.J. Lambourne, A. Redensek, C.A. Albers, V. Amstislavskiy, S. Ashford, K. Berentsen, L. Bomba, G. Bourque, D. Bujold, S. Busche, M. Caron, S.-H. Chen, W. Cheung, O. Delaneau, E.T. Dermitzakis, H. Elding, I. Colgiu, F.O. Bagger, P. Flicek, E. Habibi, V. Iotchkova, E. Janssen-Megens, B. Kim, H. Lehrach, E. Lowy, A. Mandoli, F. Matarese, M.T. Maurano, J.A. Morris, V. Pancaldi, F. Pourfarzad, K. Rehnstrom, A. Rendon, T. Risch, N. Sharifi, M.-M. Simon, M. Sultan, A. Valencia, K. Walter, S.-Y. Wang, M. Frontini, S.E. Antonarakis, L. Clarke, M.-L. Yaspo, S. Beck, R. Guigo, D. Rico, J.H.A. Martens, W.H. Ouwehand, T.W. Kuijpers, D.S. Paul, H.G. Stunnenberg, O. Stegle, K. Downes, T. Pastinen, N. Soranzo, Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells, Cell. 167 (2016) 1398–1414.e24.
[109] G.R. Keele, B.C. Quach, J.W. Israel, G.A. Chappell, L. Lewis, A. Safi, J.M. Simon, P. Cotney, G.E. Crawford, W. Valdar, I. Rusyn, T.S. Furey, Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation, PLoS Genet. 16 (2020) e1008537.
[110] N. Kumasaka, A.J. Knights, D.J. Gaffney, High-resolution genetic mapping of putative causal interactions between regions of open chromatin, Nat. Genet. 51 (2019) 128–137.
[111] B.E. Mittleman, S. Pott, S. Warland, T. Zeng, Z. Mu, M. Kaur, Y. Gilad, Y. Li, Alternative polyadenylation mediates genetic regulation of gene expression, Elife. 9 (2020). https://doi.org/10.7554/eLife.57492.
[112] O.K. Yoon, T.Y. Hsu, J.H. Im, R.B. Brem, Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells, PLoS Genet. 8 (2012) e1002882.
[113] N.E. Banovich, X. Lan, G. McVicker, B. van de Geijn, J.F. Degner, J.D. Blischak, J. Roux, J.K. Pritchard, Y. Gilad, Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels, PLoS Genet. 10 (2014) e1004663.
[114] S. Das, L. Forer, S. Schönherr, C. Sidore, A.E. Locke, A. Kwong, S.I. Vrieze, E.Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W.G. Iacono, A. Swaroop, L.J. Scott, F. Cucca, F. Kronenberg, M. Boehnke,

G.R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods, Nat. Genet. 48 (2016) 1284–1287.
[115] genimpute: Portable genotype imputation pipeline used by the eQTL Catalogue, Github, n.d. https://github.com/eQTL-Catalogue/genimpute (accessed April 27, 2023).
[116] rnaseq: RNA-seq quantification pipeline used by the eQTL Catalogue, Github, n.d. https://github.com/eQTL-Catalogue/rnaseq (accessed May 1, 2023).
[117] qcnorm: This pipeline will run QC measures of both genotype and phenotype data and will normalise quantified phenotypes, Github, n.d. https://github.com/eQTL-Catalogue/qcnorm (accessed May 1, 2023).
[118] qtlmap: Portable eQTL analysis and statistical fine mapping workflow used by the eQTL Catalogue, Github, n.d. https://github.com/eQTL-Catalogue/qtlmap (accessed May 1, 2023).
[119] coverage_plot: This repo contains a nextflow pipeline to generate the coverage plots for QTL interpretation, Github, n.d. https://github.com/kerimoff/coverage_plot (accessed May 3, 2023).
[120] colocalisation: This repo contains a nextflow pipeline to do colocalisation analysis of QTLs against GWASs, Github, n.d. https://github.com/eQTL-Catalogue/colocalisation (accessed May 3, 2023).
[121] R. Stark, M. Grzelak, J. Hadfield, RNA sequencing: the teenage years, Nat. Rev. Genet. 20 (2019) 631–656.
[122] S. Marguerat, J. Bähler, RNA-seq: from technology to biology, Cell. Mol. Life Sci. 67 (2010) 569–579.
[123] D. Sarantopoulou, T.G. Brooks, S. Nayak, A. Mrčela, N.F. Lahens, G.R. Grant, Comparative evaluation of full-length isoform quantification from RNA-Seq, BMC Bioinformatics. 22 (2021) 266.
[124] D.C. Wu, J. Yao, K.S. Ho, A.M. Lambowitz, C.O. Wilke, Limitations of alignment-free tools in total RNA-seq quantification, BMC Genomics. 19 (2018) 510.
[125] R. Chandramohan, P.-Y. Wu, J.H. Phan, M.D. Wang, Benchmarking RNA-Seq quantification tools, Conf. Proc. IEEE Eng. Med. Biol. Soc. 2013 (2013) 647–650.
[126] M. Teng, M.I. Love, C.A. Davis, S. Djebali, A. Dobin, B.R. Graveley, S. Li, C.E. Mason, S. Olson, D. Pervouchine, C.A. Sloan, X. Wei, L. Zhan, R.A. Irizarry, A benchmark for RNA-seq quantification pipelines, Genome Biol. 17 (2016) 74.
[127] N. Kerimov, R. Tambets, J.D. Hayhurst, I. Rahu, P. Kolberg, U. Raudvere, I. Kuzmin, A. Chowdhary, A. Vija, H.J. Teras, M. Kanai, J. Ulirsch, M. Ryten, J. Hardy, S. Guelfi, D. Trabzuni, S. Kim-Hellmuth, W. Rayner, H. Finucane, H. Peterson, A. Mosaku, H. Parkinson, K. Alasoo, eQTL Catalogue 2023: New datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs, PLoS Genet. 19 (2023) e1010932.
[128] S. Schaarschmidt, A. Fischer, E. Zuther, D.K. Hincha, Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana, Int. J. Mol. Sci. 21 (2020). https://doi.org/10.3390/ijms21051720.
[129] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics. 29 (2013) 15–21.

[130] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods. 12 (2015) 357–360.
[131] C. Angelini, D. De Canditiis, I. De Feis, Computational approaches for isoform detection and estimation: good and bad news, BMC Bioinformatics. 15 (2014) 135.
[132] N.L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (2016) 525–527.
[133] R. Patro, G. Duggal, M.I. Love, R.A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression, Nat. Methods. 14 (2017) 417–419.
[134] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics. 30 (2014) 923–930.
[135] J.-P. Verta, A. Jacobs, The role of alternative splicing in adaptation and evolution, Trends Ecol. Evol. 37 (2022) 299–308.
[136] S. Anders, A. Reyes, W. Huber, Detecting differential usage of exons from RNA-seq data, Genome Res. 22 (2012) 2008–2017.
[137] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics. 12 (2011) 323.
[138] R. Patro, S.M. Mount, C. Kingsford, Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms, Nat. Biotechnol. 32 (2014) 462–464.
[139] A. Srivastava, H. Sarkar, N. Gupta, R. Patro, RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes, Bioinformatics. 32 (2016) i192–i200.
[140] C. Soneson, M.I. Love, R. Patro, S. Hussain, D. Malhotra, M.D. Robinson, A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs, Life Sci Alliance. 2 (2019). https://doi.org/10.26508/lsa.201800175.
[141] H. Ongen, E.T. Dermitzakis, Alternative Splicing QTLs in European and African Populations, Am. J. Hum. Genet. 97 (2015) 567–575.
[142] D.R. Zerbino, P. Achuthan, W. Akanni, M.R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C.G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O.G. Izuogu, S.H. Janacek, T. Juettemann, J.K. To, M.R. Laird, I. Lavidas, Z. Liu, J.E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D.N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C.K. Ong, A. Parker, M. Patricio, H.S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S.E. Hunt, M. Kostadima, N. Langridge, F.J. Martin, M. Muffato, E. Perry, M. Ruffier, D.M. Staines, S.J. Trevanion, B.L. Aken, F. Cunningham, A. Yates, P. Flicek, Ensembl 2018, Nucleic Acids Res. 46 (2018) D754–D761.
[143] Y. Katz, E.T. Wang, E.M. Airoldi, C.B. Burge, Analysis and design of RNA sequencing experiments for identifying isoform regulation, Nat. Methods. 7 (2010) 1009–1015.
[144] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, Y. Hayashizaki, Cap analysis gene expression for high-throughput analysis of transcriptional starting point and

identification of promoter usage, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 15776–15781.

[145] FANTOM Consortium and the RIKEN PMI and CLST (DGT), A.R.R. Forrest, H. Kawaji, M. Rehli, J.K. Baillie, M.J.L. de Hoon, V. Haberle, T. Lassmann, I.V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C.J. Mungall, T.F. Meehan, S. Schmeier, N. Bertin, M. Jørgensen, E. Dimont, E. Arner, C. Schmidl, U. Schaefer, Y.A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C.A. Semple, Y. Ishizu, R.S. Young, M. Francescatto, I. Alam, D. Albanese, G.M. Altschuler, T. Arakawa, J.A.C. Archer, P. Arner, M. Babina, S. Rennie, P.J. Balwierz, A.G. Beckhouse, S. Pradhan-Bhatt, J.A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A.M. Burroughs, A. Califano, C.V. Cannistraci, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H.C. Clevers, E. Dalla, C.A. Davis, M. Detmar, A.D. Diehl, T. Dohi, F. Drabløs, A.S.B. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M.C. Farach-Carson, G.J. Faulkner, A.V. Favorov, M.E. Fisher, M.C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furino, J.-I. Furusawa, T.B. Geijtenbeek, A.P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T.J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K.J. Hitchens, S.J. Ho Sui, O.M. Hofmann, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B.R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A.S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y.I. Kawamura, T. Kawashima, J.S. Kempfle, T.J. Kenna, J. Kere, L.M. Khachigian, T. Kitamura, S.P. Klinken, A.J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A.T. Kwon, J.F.J. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R.-I. Manabe, J.C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D.A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C.L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, S. Noma, T. Noazaki, S. Ogishima, N. Ohkura, H. Ohimiya, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D.A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J.G.D. Prendergast, O.J.L. Rackham, J.A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M.B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, S. Savvi, A. Saxena, C. Schneider, E.A. Schultes, G.G. Schulze-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J.W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R.K. Swoboda, P.A.C. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyodo, T. Toyoda, E. Valen, M. van de Wetering, L.M. van den Berg, R. Verado, D. Vijayan, I.E. Vorontsov, W.W. Wasserman, S. Watanabe, C.A. Wells, L.N. Winteringham, E. Wolvetang, E.J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S.E. Zabierowski, P.G. Zhang, X. Zhao, S. Zucchelli, K.M. Summers, H. Suzuki, C.O. Daub, J. Kawai, P. Heutink, W. Hide, T.C. Freeman, B. Lenhard, V.B. Bajic, M.S. Taylor, V.J. Makeev, A. Sandelin, D.A. Hume, P. Carninci, Y. Hayashizaki, A promoter-level mammalian expression atlas, Nature. 507 (2014) 462–470.

[146] A. Vija, K. Alasoo, Improved detection of genetic effects on promoter usage with augmented transcript annotations, bioRxiv. (2022) 2022.07.12.499800. https://doi.org/10.1101/2022.07.12.499800.
[147] C. Trapnell, D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq, Nat. Biotechnol. 31 (2013) 46–53.
[148] N. Leng, J.A. Dawson, J.A. Thomson, V. Ruotti, A.I. Rissman, B.M.G. Smits, J.D. Haag, M.N. Gould, R.M. Stewart, C. Kendziorski, EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments, Bioinformatics. 29 (2013) 1035–1043.
[149] T. Steijger, J.F. Abril, P.G. Engström, F. Kokocinski, RGASP Consortium, T.J. Hubbard, R. Guigó, J. Harrow, P. Bertone, Assessment of transcript reconstruction methods for RNA-seq, Nat. Methods. 10 (2013) 1177–1184.
[150] Y.I. Li, D.A. Knowles, J. Humphrey, A.N. Barbeira, S.P. Dickinson, H.K. Im, J.K. Pritchard, Annotation-free quantification of RNA splicing using LeafCutter, Nat. Genet. 50 (2018) 151–158.
[151] N. Kerimov, Töökindla ja teisaldatava töövoo väljatöötamine molekulaarsete tunnustega seotud geneetiliste variantide tuvastamiseks mitmetest andmestikest, Master's, University of Tartu, 2019. https://dspace.ut.ee/handle/10062/66408 (accessed April 14, 2023).
[152] K. Alasoo, J. Rodrigues, S. Mukhopadhyay, A.J. Knights, A.L. Mann, K. Kundu, HIPSCI Consortium, C. Hale, G. Dougan, D.J. Gaffney, Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response, Nat. Genet. 50 (2018) 424–431.
[153] H.J. Zhou, L. Li, Y. Li, W. Li, J.J. Li, PCA outperforms popular hidden variable inference methods for molecular QTL mapping, Genome Biol. 23 (2022) 210.
[154] J.T. Leek, J.D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, PLoS Genet. 3 (2007) 1724–1735.
[155] O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses, Nat. Protoc. 7 (2012) 500–507.
[156] O. Stegle, L. Parts, R. Durbin, J. Winn, A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies, PLoS Comput. Biol. 6 (2010) e1000770.
[157] S. Mostafavi, A. Battle, X. Zhu, A.E. Urban, D. Levinson, S.B. Montgomery, D. Koller, Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge, PLoS One. 8 (2013) e68141.
[158] J.R. Davis, L. Fresard, D.A. Knowles, M. Pala, C.D. Bustamante, A. Battle, S.B. Montgomery, An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants, Am. J. Hum. Genet. 98 (2016) 216–224.
[159] H. Ongen, A. Buil, A.A. Brown, E.T. Dermitzakis, O. Delaneau, Fast and efficient QTL mapper for thousands of molecular phenotypes, Bioinformatics. 32 (2016) 1479–1485.
[160] H. Li, Tabix: fast retrieval of sequence features from generic TAB-delimited files, Bioinformatics. 27 (2011) 718–719.
[161] M.S. Kris A. Wetterstrand, The Cost of Sequencing a Human Genome, Genome.gov. (2019). https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (accessed April 28, 2023).

[162] Whole Genome Sequencing, Dante Labs World. (n.d.). https://www.dantelabs.com/ (accessed April 28, 2023).
[163] E. Mullin, The Era of Fast, Cheap Genome Sequencing Is Here, Wired. (2022). https://www.wired.com/story/the-era-of-fast-cheap-genome-sequencing-is-here/ (accessed April 28, 2023).
[164] A. by Y. Shin, Whole Genome Sequencing cost 2023, 3billion. (n.d.). https://3billion.io/blog/whole-genome-sequencing-cost-2023 (accessed April 17, 2023).
[165] Z.D. Stephens, S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, M.J. Efron, R. Iyer, M.C. Schatz, S. Sinha, G.E. Robinson, Big Data: Astronomical or Genomical?, PLoS Biol. 13 (2015) e1002195.
[166] P.A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M.U. Garcia, P. Di Tommaso, S. Nahnsen, The nf-core framework for community-curated bioinformatics pipelines, Nat. Biotechnol. 38 (2020) 276–278.
[167] P. Di Tommaso, M. Chatzou, E.W. Floden, P.P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows, Nat. Biotechnol. 35 (2017) 316–319.
[168] J. Leipzig, Computational Pipelines and Workflows in Bioinformatics, in: S. Ranganathan, M. Gribskov, K. Nakai, C. Schönbach (Eds.), Encyclopedia of Bioinformatics and Computational Biology, Academic Press, Oxford, 2018: pp. 1151–1162.
[169] J. Leipzig, A review of bioinformatic pipeline frameworks, Brief. Bioinform. 18 (2017) 530–536.
[170] S.I. Feldman, Make: A Program for Maintaining Computer Programs, Bell Telephone Laboratories, 1975.
[171] J. Goecks, A. Nekrutenko, J. Taylor, Galaxy Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol. 11 (2010) R86.
[172] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M.P. Balcazar Vargas, S. Sufi, C. Goble, The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud, Nucleic Acids Res. 41 (2013) W557–61.
[173] G. Rakocevic, V. Semenyuk, W.-P. Lee, J. Spencer, J. Browning, I.J. Johnson, V. Arsenijevic, J. Nadj, K. Ghose, M.C. Suciu, S.-G. Ji, G. Demir, L. Li, B.Ç. Toptaş, A. Dolgoborodov, B. Pollex, I. Spulber, I. Glotova, P. Kómár, A.L. Stachyra, Y. Li, M. Popovic, M. Källberg, A. Jain, D. Kural, Fast and accurate genomic analyses using genome graphs, Nat. Genet. 51 (2019) 354–362.
[174] DNAnexus®, (n.d.). https://www.dnanexus.com/ (accessed May 3, 2023).
[175] B. Fjukstad, L.A. Bongo, A Review of Scalable Bioinformatics Pipelines, Data Science and Engineering. 2 (2017) 245–251.
[176] R. Ferreira da Silva, R. Filgueira, I. Pietri, M. Jiang, R. Sakellariou, E. Deelman, A characterization of workflow management systems for extreme-scale applications, Future Gener. Comput. Syst. 75 (2017) 228–238.
[177] I. Foster, Y. Zhao, I. Raicu, S. Lu, Cloud Computing and Grid Computing 360-Degree Compared, in: 2008 Grid Computing Environments Workshop, 2008: pp. 1–10.

[178] B.R. Kandukuri, R.P. V., A. Rakshit, Cloud Security Issues, in: 2009 IEEE International Conference on Services Computing, 2009: pp. 517–520.
[179] M.D. Wilkinson, M. Dumontier, I.J.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci Data. 3 (2016) 160018.
[180] G.A. Van der Auwera, M.O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K.V. Garimella, D. Altshuler, S. Gabriel, M.A. DePristo, From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline, Curr. Protoc. Bioinformatics. 43 (2013) 11.10.1–11.10.33.
[181] R.D. Peng, Reproducible research in computational science, Science. 334 (2011) 1226–1227.
[182] F. Strozzi, R. Janssen, R. Wurmus, M.R. Crusoe, G. Githinji, P. Di Tommaso, D. Belhachemi, S. Möller, G. Smant, J. de Ligt, P. Prins, Scalable Workflows and Reproducible Data Analysis for Genomics, in: M. Anisimova (Ed.), Evolutionary Genomics: Statistical and Computational Methods, Springer New York, New York, NY, 2019: pp. 723–745.
[183] V. Stodden, J. Borwein, D.H. Bailey, Setting the default to reproducible, Computational Science Research. SIAM News. 46 (2013) 4–6.
[184] V. Stodden, J. Seiler, Z. Ma, An empirical analysis of journal policy effectiveness for computational reproducibility, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) 2584–2589.
[185] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, J. Myers, Examining the Challenges of Scientific Workflows, Computer . 40 (2007) 24–32.
[186] S. Kanwal, F.Z. Khan, A. Lonie, R.O. Sinnott, Investigating reproducibility and tracking provenance - A genomic workflow case study, BMC Bioinformatics. 18 (2017) 337.
[187] N. Kulkarni, L. Alessandrì, R. Panero, M. Arigoni, M. Olivero, G. Ferrero, F. Cordero, M. Beccuti, R.A. Calogero, Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines, BMC Bioinformatics. 19 (2018) 349.
[188] G.K. Sandve, A. Nekrutenko, J. Taylor, E. Hovig, Ten simple rules for reproducible computational research, PLoS Comput. Biol. 9 (2013) e1003285.
[189] B.A. Nosek, G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, S. Buck, C.D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D.P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T.A. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E.L. Paluck, U. Simonsohn, C. Soderberg, B.A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E.J. Wagenmakers, R.

Wilson, T. Yarkoni, SCIENTIFIC STANDARDS. Promoting an open research culture, Science. 348 (2015) 1422–1425.
[190] S. Mangul, T. Mosqueiro, D. Duong, K. Mitchell, V. Sarwal, B. Hill, J. Brito, R. Littman, B. Statz, A. Lam, G. Dayama, L. Grieneisen, L. Martin, J. Flint, E. Eskin, R. Blekhman, A comprehensive analysis of the usability and archival stability of omics computational tools and resources, bioRxiv. (2018) 452532. https://doi.org/10.1101/452532.
[191] B. Grüning, J. Chilton, J. Köster, R. Dale, N. Soranzo, M. van den Beek, J. Goecks, R. Backofen, A. Nekrutenko, J. Taylor, Practical Computational Reproducibility in the Life Sciences, Cell Syst. 6 (2018) 631–635.
[192] F. Denk, Don't let useful data go to waste, Nature. 543 (2017) 7.
[193] J. Kaye, From single biobanks to international networks: developing e-governance, Hum. Genet. 130 (2011) 377–382.
[194] B.M. Knoppers, Framework for responsible sharing of genomic and health-related data, Hugo J. 8 (2014) 3.
[195] A.E. Ahmed, J.M. Allen, T. Bhat, P. Burra, C.E. Fliege, S.N. Hart, J.R. Heldenbrand, M.E. Hudson, D.D. Istanto, M.T. Kalmbach, G.D. Kapraun, K.I. Kendig, M.C. Kendzior, E.W. Klee, N. Mattson, C.A. Ross, S.M. Sharif, R. Venkatakrishnan, F.M. Fadlelmola, L.S. Mainzer, Design considerations for workflow management systems use in production genomics research and the clinic, Sci. Rep. 11 (2021) 21680.
[196] A.B. Yoo, M.A. Jette, M. Grondona, SLURM: Simple Linux Utility for Resource Management, in: Job Scheduling Strategies for Parallel Processing, Springer Berlin Heidelberg, 2003: pp. 44–60.
[197] IBM Spectrum LSF Suites, (n.d.). https://www.ibm.com/products/hpc-workload-management (accessed May 3, 2023).
[198] MOAB HPC SUITE, Adaptive Computing. (2021). https://adaptivecomputing.com/moab-hpc-suite/ (accessed May 3, 2023).
[199] W. Gentzsch, Sun Grid Engine: towards creating a compute power grid, in: Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid, 2001: pp. 35–36.
[200] N. Capit, G. Da Costa, Y. Georgiou, G. Huard, C. Martin, G. Mounie, P. Neyron, O. Richard, A batch scheduler with high level components, in: CCGrid 2005. IEEE International Symposium on Cluster Computing and the Grid, 2005., 2005: pp. 776–783 Vol. 2.
[201] D.H. Ahn, N. Bass, A. Chu, J. Garlick, M. Grondona, S. Herbein, H.I. Ingólfsson, J. Koning, T. Patki, T.R.W. Scogland, B. Springmeyer, M. Taufer, Flux: Overcoming scheduling challenges for exascale workflows, Future Gener. Comput. Syst. 110 (2020) 202–213.
[202] B. Langmead, A. Nellore, Cloud computing for genomic data analysis and collaboration, Nat. Rev. Genet. 19 (2018) 325.
[203] A. Roy, Y. Diao, U. Evani, A. Abhyankar, C. Howarth, R. Le Priol, T. Bloom, Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study, in: Proceedings of the 2017 ACM International Conference on Management of Data, Association for Computing Machinery, New York, NY, USA, 2017: pp. 187–202.
[204] Conda — Conda documentation, (n.d.). https://conda.io/en/latest/ (accessed March 26, 2019).

[205] B. Grüning, R. Dale, A. Sjödin, B.A. Chapman, J. Rowe, C.H. Tomkins-Tinch, R. Valieris, J. Köster, Bioconda Team, Bioconda: sustainable and comprehensive software distribution for the life sciences, Nat. Methods. 15 (2018) 475–476.
[206] Oracle VM VirtualBox, (n.d.). https://www.virtualbox.org/ (accessed May 3, 2023).
[207] Introducing VMware Cross-Cloud Services, VMware. (2023). https://www.vmware.com/ (accessed May 3, 2023).
[208] P. Di Tommaso, E. Palumbo, M. Chatzou, P. Prieto, M.L. Heuer, C. Notredame, The impact of Docker containers on the performance of genomic pipelines, PeerJ. 3 (2015) e1273.
[209] W.L. Schulz, T.J.S. Durant, A.J. Siddon, R. Torres, Use of application containers and workflows for genomic data analysis, J. Pathol. Inform. 7 (2016) 53.
[210] D. Merkel, Docker: Lightweight Linux Containers for Consistent Development and Deployment, Linux J. 2014 (2014). http://dl.acm.org/citation.cfm?id=2600239.2600241.
[211] G.M. Kurtzer, V. Sochat, M.W. Bauer, Singularity: Scientific containers for mobility of compute, PLoS One. 12 (2017) e0177459.
[212] W. Felter, A. Ferreira, R. Rajamony, J. Rubio, An updated performance comparison of virtual machines and Linux containers, in: 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2015: pp. 171–172.
[213] C. Boettiger, An introduction to Docker for reproducible research, Oper. Syst. Rev. 49 (2015) 71–79.
[214] S. Stolberg, Enabling Agile Testing through Continuous Integration, in: 2009 Agile Conference, 2009: pp. 369–374.
[215] B. Vasilescu, Y. Yu, H. Wang, P. Devanbu, V. Filkov, Quality and productivity outcomes relating to continuous integration in GitHub, in: Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, Association for Computing Machinery, New York, NY, USA, 2015: pp. 805–816.
[216] R.C. Martin, The open-closed principle, More C++ Gems. 19 (1996) 9.
[217] G.V. Roshchupkin, H.H.H. Adams, M.W. Vernooij, A. Hofman, C.M. Van Duijn, M.A. Ikram, W.J. Niessen, HASE: Framework for efficient high-dimensional association analyses, Sci. Rep. 6 (2016) 36076.
[218] P.M. Visscher, N.R. Wray, Q. Zhang, P. Sklar, M.I. McCarthy, M.A. Brown, J. Yang, 10 Years of GWAS Discovery: Biology, Function, and Translation, Am. J. Hum. Genet. 101 (2017) 5–22.
[219] GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues, Science. 369 (2020) 1318–1330.
[220] T. Berisa, J.K. Pickrell, Approximately independent linkage disequilibrium blocks in human populations, Bioinformatics. 32 (2016) 283–285.
[221] B. van de Geijn, G. McVicker, Y. Gilad, J.K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery, Nat. Methods. 12 (2015) 1061–1063.
[222] K. Alasoo, wiggleplotr: A small R package to make sequencing read coverage plots in R, Github, n.d. https://github.com/kauralasoo/wiggleplotr (accessed May 1, 2023).
[223] J. Morales, S. Pujar, J.E. Loveland, A. Astashyn, R. Bennett, A. Berry, E. Cox, C. Davidson, O. Ermolaeva, C.M. Farrell, R. Fatima, L. Gil, T. Goldfarb, J.M. Gonzalez, D. Haddad, M. Hardy, T. Hunt, J. Jackson, V.S. Joardar, M. Kay,

V.K. Kodali, K.M. McGarvey, A. McMahon, J.M. Mudge, D.N. Murphy, M.R. Murphy, B. Rajput, S.H. Rangwala, L.D. Riddick, F. Thibaud-Nissen, G. Threadgold, A.R. Vatsan, C. Wallin, D. Webb, P. Flicek, E. Birney, K.D. Pruitt, A. Frankish, F. Cunningham, T.D. Murphy, A joint NCBI and EMBL-EBI transcript set for clinical genomics and research, Nature. 604 (2022) 310–315.

[224] M.I. Kurki, J. Karjalainen, P. Palta, T.P. Sipilä, K. Kristiansson, K.M. Donner, M.P. Reeve, H. Laivuori, M. Aavikko, M.A. Kaunisto, A. Loukola, E. Lahtela, H. Mattsson, P. Laiho, P. Della Briotta Parolo, A.A. Lehisto, M. Kanai, N. Mars, J. Rämö, T. Kiiskinen, H.O. Heyne, K. Veerapen, S. Rüeger, S. Lemmelä, W. Zhou, S. Ruotsalainen, K. Pärn, T. Hiekkalinna, S. Koskelainen, T. Paajanen, V. Llorens, J. Gracia-Tabuenca, H. Siirtola, K. Reis, A.G. Elnahas, B. Sun, C.N. Foley, K. Aalto-Setälä, K. Alasoo, M. Arvas, K. Auro, S. Biswas, A. Bizaki-Vallaskangas, O. Carpen, C.-Y. Chen, O.A. Dada, Z. Ding, M.G. Ehm, K. Eklund, M. Färkkilä, H. Finucane, A. Ganna, A. Ghazal, R.R. Graham, E.M. Green, A. Hakanen, M. Hautalahti, Å.K. Hedman, M. Hiltunen, R. Hinttala, I. Hovatta, X. Hu, A. Huertas-Vazquez, L. Huilaja, J. Hunkapiller, H. Jacob, J.-N. Jensen, H. Joensuu, S. John, V. Julkunen, M. Jung, J. Junttila, K. Kaarniranta, M. Kähönen, R. Kajanne, L. Kallio, R. Kälviäinen, J. Kaprio, FinnGen, N. Kerimov, J. Kettunen, E. Kilpeläinen, T. Kilpi, K. Klinger, V.-M. Kosma, T. Kuopio, V. Kurra, T. Laisk, J. Laukkanen, N. Lawless, A. Liu, S. Longerich, R. Mägi, J. Mäkelä, A. Mäkitie, A. Malarstig, A. Mannermaa, J. Maranville, A. Matakidou, T. Meretoja, S.V. Mozaffari, M.E.K. Niemi, M. Niemi, T. Niiranen, C.J. O Donnell, M.E. Obeidat, G. Okafo, H.M. Ollila, A. Palomäki, T. Palotie, J. Partanen, D.S. Paul, M. Pelkonen, R.K. Pendergrass, S. Petrovski, A. Pitkäranta, A. Platt, D. Pulford, E. Punkka, P. Pussinen, N. Raghavan, F. Rahimov, D. Rajpal, N.A. Renaud, B. Riley-Gillis, R. Rodosthenous, E. Saarentaus, A. Salminen, E. Salminen, V. Salomaa, J. Schleutker, R. Serpi, H.-Y. Shen, R. Siegel, K. Silander, S. Siltanen, S. Soini, H. Soininen, J.H. Sul, I. Tachmazidou, K. Tasanen, P. Tienari, S. Toppila-Salmi, T. Tukiainen, T. Tuomi, J.A. Turunen, J.C. Ulirsch, F. Vaura, P. Virolainen, J. Waring, D. Waterworth, R. Yang, M. Nelis, A. Reigo, A. Metspalu, L. Milani, T. Esko, C. Fox, A.S. Havulinna, M. Perola, S. Ripatti, A. Jalanko, T. Laitinen, T.P. Mäkelä, R. Plenge, M. McCarthy, H. Runz, M.J. Daly, A. Palotie, FinnGen provides genetic insights from a well-phenotyped isolated population, Nature. 613 (2023) 508–518.

[225] L. Kolberg, N. Kerimov, H. Peterson, K. Alasoo, Co-expression analysis reveals interpretable gene modules controlled by trans-acting genetic variants, Elife. 9 (2020). https://doi.org/10.7554/eLife.58705.

[226] H. Hautakangas, B.S. Winsvold, S.E. Ruotsalainen, G. Bjornsdottir, A.V.E. Harder, L.J.A. Kogelman, L.F. Thomas, R. Noordam, C. Benner, P. Gormley, V. Artto, K. Banasik, A. Bjornsdottir, D.I. Boomsma, B.M. Brumpton, K.S. Burgdorf, J.E. Buring, M.A. Chalmer, I. de Boer, M. Dichgans, C. Erikstrup, M. Färkkilä, M.E. Garbrielsen, M. Ghanbari, K. Hagen, P. Häppölä, J.-J. Hottenga, M.G. Hrafnsdottir, K. Hveem, M.B. Johnsen, M. Kähönen, E.S. Kristoffersen, T. Kurth, T. Lehtimäki, L. Lighart, S.H. Magnusson, R. Malik, O.B. Pedersen, N. Pelzer, B.W.J.H. Penninx, C. Ran, P.M. Ridker, F.R. Rosendaal, G.R. Sigurdardottir, A.H. Skogholt, O.A. Sveinsson, T.E. Thorgeirsson, H. Ullum, L.S. Vijfhuizen, E. Widén, K.W. van Dijk, International Headache Genetics Consortium, HUNT All-in Headache, Danish Blood Donor Study Genomic Cohort, A. Aromaa, A.C. Belin, T. Freilinger, M.A. Ikram, M.-R. Järvelin, O.T.

Raitakari, G.M. Terwindt, M. Kallela, M. Wessman, J. Olesen, D.I. Chasman, D.R. Nyholt, H. Stefánsson, K. Stefansson, A.M.J.M. van den Maagdenberg, T.F. Hansen, S. Ripatti, J.-A. Zwart, A. Palotie, M. Pirinen, Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles, Nat. Genet. 54 (2022) 152–160.

[227] J.T. Rämö, T. Kiiskinen, R. Seist, K. Krebs, M. Kanai, J. Karjalainen, M. Kurki, E. Hämäläinen, P. Häppölä, A.S. Havulinna, H. Hautakangas, FinnGen, R. Mägi, P. Palta, T. Esko, A. Metspalu, M. Pirinen, K.J. Karczewski, S. Ripatti, L. Milani, K.M. Stankovic, A. Mäkitie, M.J. Daly, A. Palotie, Genome-wide screen of otosclerosis in population biobanks: 27 loci and shared associations with skeletal structure, Nat. Commun. 14 (2023) 157.

[228] B. Khatri, K.L. Tessneer, A. Rasmussen, F. Aghakhanian, T.R. Reksten, A. Adler, I. Alevizos, J.-M. Anaya, L.A. Aqrawi, E. Baecklund, J.G. Brun, S.M. Bucher, M.-L. Eloranta, F. Engelke, H. Forsblad-d'Elia, S.B. Glenn, D. Hammenfors, J. Imgenberg-Kreuz, J.L. Jensen, S.J.A. Johnsen, M.V. Jonsson, M. Kvarnström, J.A. Kelly, H. Li, T. Mandl, J. Martín, G. Nocturne, K.B. Norheim, Ø. Palm, K. Skarstein, A.M. Stolarczyk, K.E. Taylor, M. Teruel, E. Theander, S. Venuturupalli, D.J. Wallace, K.M. Grundahl, K.S. Hefner, L. Radfar, D.M. Lewis, D.U. Stone, C.E. Kaufman, M.T. Brennan, J.M. Guthridge, J.A. James, R.H. Scofield, P.M. Gaffney, L.A. Criswell, R. Jonsson, P. Eriksson, S.J. Bowman, R. Omdal, L. Rönnblom, B. Warner, M. Rischmueller, T. Witte, A.D. Farris, X. Mariette, M.E. Alarcon-Riquelme, PRECISESADS Clinical Consortium, C.H. Shiboski, Sjögren's International Collaborative Clinical Alliance (SICCA), M. Wahren-Herlenius, W.-F. Ng, UK Primary Sjögren's Syndrome Registry, K.L. Sivils, I. Adrianto, G. Nordmark, C.J. Lessard, Genome-wide association study identifies Sjögren's risk loci with functional implications in immune and glandular cells, Nat. Commun. 13 (2022) 4287.

[229] I.E. Jansen, S.J. van der Lee, D. Gomez-Fonseca, I. de Rojas, M.C. Dalmasso, B. Grenier-Boley, A. Zettergren, A. Mishra, M. Ali, V. Andrade, C. Bellenguez, L. Kleineidam, F. Küçükali, Y.J. Sung, N. Tesí, E.M. Vromen, D.P. Wightman, D. Alcolea, M. Alegret, I. Alvarez, P. Amouyel, L. Athanasiu, S. Bahrami, H. Bailly, O. Belbin, S. Bergh, L. Bertram, G.J. Biessels, K. Blennow, R. Blesa, M. Boada, A. Boland, K. Buerger, Á. Carracedo, L. Cervera-Carles, G. Chene, J.A.H.R. Claassen, S. Debette, J.-F. Deleuze, P.P. de Deyn, J. Diehl-Schmid, S. Djurovic, O. Dols-Icardo, C. Dufouil, E. Duron, E. Düzel, EADB consortium, T. Fladby, J. Fortea, L. Frölich, P. García-González, M. Garcia-Martinez, I. Giegling, O. Goldhardt, J. Gobom, T. Grimmer, A. Haapasalo, H. Hampel, O. Hanon, L. Hausner, S. Heilmann-Heimbach, S. Helisalmi, M.T. Heneka, I. Hernández, S.-K. Herukka, H. Holstege, J. Jarholm, S. Kern, A.-B. Knapskog, A.M. Koivisto, J. Kornhuber, T. Kuulasmaa, C. Lage, C. Laske, V. Leinonen, P. Lewczuk, A. Lleó, A.L. de Munain, S. Lopez-Garcia, W. Maier, M. Marquié, M.O. Mol, L. Montrreal, F. Moreno, S. Moreno-Grau, G. Nicolas, M.M. Nöthen, A. Orellana, L. Pålhaugen, J.M. Papma, F. Pasquier, R. Perneczky, O. Peters, Y.A.L. Pijnenburg, J. Popp, D. Posthuma, A. Pozueta, J. Priller, R. Puerta, I. Quintela, I. Ramakers, E. Rodriguez-Rodriguez, D. Rujescu, I. Saltvedt, P. Sanchez-Juan, P. Scheltens, N. Scherbaum, M. Schmid, A. Schneider, G. Selbæk, P. Selnes, A. Shadrin, I. Skoog, H. Soininen, L. Tárraga, S. Teipel, GR@ACE study group, B. Tijms, M. Tsolaki, C. Van Broeckhoven, J. Van Dongen, J.C. van Swieten, R. Vandenberghe, J.-S. Vidal, P.J. Visser, J.

Vogelgsang, M. Waern, M. Wagner, J. Wiltfang, M.M.J. Wittens, H. Zetterberg, M. Zulaica, C.M. van Duijn, M. Bjerke, S. Engelborghs, F. Jessen, C.E. Teunissen, P. Pastor, M. Hiltunen, M. Ingelsson, O.A. Andreassen, J. Clarimón, K. Sleegers, A. Ruiz, A. Ramirez, C. Cruchaga, J.-C. Lambert, W. van der Flier, Genome-wide meta-analysis for Alzheimer's disease cerebrospinal fluid biomarkers, Acta Neuropathol. 144 (2022) 821–842.

[230] P. Sole-Navais, J. Juodakis, K. Ytterberg, X. Wu, M. Vaudel, Ø. Helgeland, C. Flatley, F. Geller, P. Magnus, O.A. Andreassen, P. Njølstad, B. Feenstra, L.J. Muglia, S. Johansson, G. Zhang, B. Jacobsson, Genome-wide analyses of neonatal jaundice reveal a marked departure from adult bilirubin metabolism, bioRxiv. (2022). https://doi.org/10.1101/2022.12.14.22283348.

[231] N. Pujol-Gualdo, K. Läll, M. Lepamets, Estonian Biobank Research Team, H.-R. Rossi, R.K. Arffman, T.T. Piltonen, R. Mägi, T. Laisk, Advancing our understanding of genetic risk factors and potential personalized strategies for pelvic organ prolapse, Nat. Commun. 13 (2022) 3584.

[232] M. Koel, U. Võsa, M. Jõeloo, K. Läll, N.P. Gualdo, H. Laivuori, S. Lemmelä, Estonian Biobank Research Team, FinnGen, M. Daly, P. Palta, R. Mägi, T. Laisk, GWAS meta-analyses clarify genetics of cervical phenotypes and inform risk stratification for cervical cancer, Hum. Mol. Genet. (2023). https://doi.org/10.1093/hmg/ddad043.

[233] K. Tsuo, W. Zhou, Y. Wang, M. Kanai, S. Namba, R. Gupta, L. Majara, L.L. Nkambule, T. Morisaki, Y. Okada, B.M. Neale, Global Biobank Meta-analysis Initiative, M.J. Daly, A.R. Martin, Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity, Cell Genom. 2 (2022) 100212.

[234] W.R. Reay, M.P. Geaghan, 23andMe Research Team, M.J. Cairns, The genetic architecture of pneumonia susceptibility implicates mucin biology and a relationship with psychiatric illness, Nat. Commun. 13 (2022) 3756.

[235] E.C. Saarentaus, J. Karjalainen, J.T. Rämö, T. Kiiskinen, A.S. Havulinna, J. Mehtonen, H. Hautakangas, S. Ruotsalainen, M. Tamlander, N. Mars, FINNGEN, S. Toppila-Salmi, M. Pirinen, M. Kurki, S. Ripatti, M. Daly, T. Palotie, A. Mäkitie, A. Palotie, Inflammatory and infectious upper respiratory diseases associate with 41 genomic loci and type 2 inflammation, Nat. Commun. 14 (2023) 83.

[236] M. Oliva, K. Demanelis, Y. Lu, M. Chernoff, F. Jasmine, H. Ahsan, M.G. Kibriya, L.S. Chen, B.L. Pierce, DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits, Nat. Genet. 55 (2023) 112–122.

[237] K.M. Chen, A.K. Wong, O.G. Troyanskaya, J. Zhou, A sequence-based global map of regulatory activity for deciphering human genetics, Nat. Genet. 54 (2022) 940–949.

[238] P.J. Castaldi, A. Abood, C.R. Farber, G.M. Sheynkman, Bridging the splicing gap in human genetics with long-read RNA sequencing: finding the protein isoform drivers of disease, Hum. Mol. Genet. 31 (2022) R123–R136.

[239] B.J. Schmiedel, C. Gonzalez-Colin, V. Fajardo, J. Rocha, A. Madrigal, C. Ramírez-Suástegui, S. Bhattacharyya, H. Simon, J.A. Greenbaum, B. Peters, G. Seumois, F. Ay, V. Chandra, P. Vijayanand, Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type-dependent effects of disease-risk variants, Sci Immunol. 7 (2022) eabm2508.

[240] J.A. Morris, C. Caragine, Z. Daniloski, J. Domingo, T. Barry, L. Lu, K. Davis, M. Ziosi, D.A. Glinos, S. Hao, E.P. Mimitou, P. Smibert, K. Roeder, E. Katsevich, T. Lappalainen, N.E. Sanjana, Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens, Science. (2023) eadh7699.

[241] F. Cunningham, J.E. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, O. Austine-Orimoloye, A.G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C.G. Giron, T. Genez, J.G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J.C. Marugán, S. Mohanan, A. Mushtaq, M. Naven, D.N. Ogeh, A. Parker, A. Parton, M. Perry, I. Piližota, I. Prosovetskaia, M.P. Sakthivel, A.I.A. Salam, B.M. Schmitt, H. Schuilenburg, D. Sheppard, J.G. Pérez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M.-M. Suner, M. Szpak, A. Thormann, F.F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T.A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S.E. Hunt, G.R. IIsley, J.E. Loveland, F.J. Martin, B. Moore, J.M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S.J. Trevanion, S. Dyer, P.W. Harrison, K.L. Howe, A.D. Yates, D.R. Zerbino, P. Flicek, Ensembl 2022, Nucleic Acids Res. 50 (2022) D988–D995.

[242] A. Kwong, A.P. Boughton, M. Wang, P. VandeHaar, M. Boehnke, G. Abecasis, H.M. Kang, FIVEx: an interactive multi-tissue eQTL browser, bioRxiv. (2021) 2021.01.22.426874. https://doi.org/10.1101/2021.01.22.426874.

ACKNOWLEDGEMENT

On a typical day during our Data Mining practice session, Dmytro Fishman (or Dima, as we fondly call him) made a passing comment about this really impressive guy who'd just returned from studying at Cambridge University. He was purportedly an expert in genetics – well, at least more knowledgeable than most folks around, and his name is **Kaur**. I had this mental image of Kaur as an eccentric, white-bearded professor, maybe around sixty. But upon googling him, he was nothing like what I'd pictured. When we met the first time, I confessed to him that I was never a fan of biology back in school and that my understanding of genetics was embarrassingly limited. However, I was deeply fascinated by the recent developments in the field as it had become increasingly computational, aligning more with my software engineering background. He said that my eagerness to learn was all that mattered, and so he agreed to supervise my master's thesis. Sometimes, in our lives, we encounter certain individuals who irrevocably change our trajectories. For me, Kaur is definitely one of those. He's exceptionally brilliant, incredibly compassionate, and a supportive mentor who stood by my side, irrespective of the decisions I made. Without his guidance, I can confidently say that I would've given up on academia in my second year, without a second thought. The fact that our paths crossed and we ended up collaborating is something I consider an extraordinary stroke of luck. This partnership turned out to be a remarkably fruitful one. To say I owe Kaur is a massive understatement – I feel a lifelong debt to him. No matter what I do, it seems like it will never be enough to repay him for the transformative impact he's had on my life. **Thank you Kaur!**

I'm also very grateful to the amazing individuals in our lab: **Ralf**, **Peep**, **Krista**, **Ida**, **Anastassia**, **Dzvinka**, and **Kristiina**. As the first student member of this team, I've absorbed a wealth of knowledge from most of you, and I sincerely hope that I was able to contribute in some way to your journey as well. I think Kaur basically eliminates annoying people in the first round and keeps the cool ones. Here's a friendly piece of advice for you: when you're in a one-on-one meeting with Kaur, and his brain starts firing off three truly amazing ideas every minute, don't hesitate to stop him. Trust me, you'll thank me for this tip later. ;)

I've heard more than a few PhD students grumbling about how their journey is filled with stress and hardship. In contrast, I'm thrilled to share that my experience has been absolutely delightful. Likely, it's due to a combination of factors: the positive “Kaur” effect, the fact that I'm a little older than a typical PhD student, and having had my fair share of experience in “non-academic” work environments. But one of the most significant contributors has undoubtedly been the wonderful folks in our institute. I owe a heap of gratitude to each and every one of them! **Uku**, with his serene nature, instilled in me the “it is OK” mindset; **the Professor** (she knows herself) who stood as a beacon of inspiration, earning the title of my favourite professor; then there's **Liis**, who not

only invented “banana-time” but is also remarkably honest when it's called for; **Priit (Lemps)** has a knack for opening my eyes to fresh perspectives; **Taavi** never shies away from an interesting discussion and embodies a “always happy to help” attitude; **Ivan** stands out for his uniqueness and might just be the nicest person I know; **Elena** has this rare ability to transfer some of her non-toxic positivity when you need some; **Diana**, **Heleri**, **Ahto**, **Erik** and **Mari-Liis**, whose genuine and candid conversations I cherish; **Jaak**, who helped me zoom out to view the bigger picture from time to time; **Meelis** and **Mark**, who consistently exemplifies how to stay composed, cool and pleasant in the challenging world of academia; and **Kaido** for our thoughtful meetings and deep-diving discussions on the topics of free-will and the role of luck in life. Indeed, this isn't an exhaustive list of those I'm grateful to, and if I've inadvertently missed someone, please accept my sincere apologies in advance.

A big shout-out to the admin team at the institute of computer science for fostering a excellent workspace, while always offering a helping hand with our mildly annoying issues in the most patient manner: **Jaak Vilo, Heisi Kurig, Piret Orav, Heili Kase, Martin Kaljula, Eeva Kilk, Natali Belinska, Jaanika Seli, Karis Meister, Ülle Holm, Maarja Kungla, Kersti Taurus, Liivi Luik, Henry Narits, Reili Liiver, Anneli Vainumäe, Käroliin Jääger, Saili Petti, Sirli Urbas, Siiri Pilt, Laura-Kristiina Rand and Daisy Alatare**.

Conducting research in genomics without a High-Performance Computer (HPC) would be unimaginable given the vast amounts of data involved. I fully understand the immense effort required to constantly enhance and sustain such a complex system. I also recognize the unique advantage we, researchers at UT, possess by having such a facility within our reach. My heartfelt appreciation goes out to the entire HPC team, particularly **Sander**, **Ulvi**, **Ott**, and **Gular**. Their quick action in resolving any issues I encounter with the HPC is truly commendable.

A very special thanks to my internal reviewer **Sulev Reisberg**, who not only made this thesis better with his constructive comments but also did it extremely rapidly, and spent another session with me to discuss the comments which I didn’t agree with. I highly appreciate your time and effort you put in this work and I could not have asked for a better internal reviewer.

I must express my gratitude to my external opponents who generously took time out of their precious summer days to read and provide feedback on my thesis. I'm truly thankful for **Prof. Gregory C. Gibson** and **Dr. Emma Davenport**, experts in their fields, who graciously accepted the roles as my doctoral defence opponents.

I would be remiss not to express my gratitude and appreciation to my esteemed co-authors, who have played a pivotal role in the production of our published papers**. James Hajhurst**, **Daniel Zerbino**, **Ralf Tambets**, **Ivan Kuzmin**, **Katerina Peikova** and all the others – your insightful contributions, relentless dedication, and unwavering commitment have been instrumental to our collective success. Each of you brought unique perspectives and expertise

that enriched our work, fostering a truly collaborative spirit. Thank you for this exceptional partnership!

A big thank you to the brave and dedicated team at Nightingale Health, who are actively shaping the future of preventative health. It's inspiring to see their shared courage and intelligence, and I'm glad to be a part of it. I'm particularly grateful to **Jeff** for believing in me and my ability to handle both my work and my PhD thesis at the same time, and for his constant support along this path.

When you love your work, it is hard to rest. To ensure the rest it is necessary to have very interesting people around who can fill the time with joy and laughter. I have cool friends living in Tallinn, **Ehtiram** and **Miyu**, who specifically visited me in Tartu to clarify why I like Tartu so much, and don't like Tallinn. They correctly identified that I don't live in Tartu but live in a bubble of highly intelligent and nice people in Tartu, which makes Tallinn look not as good. Thank you, guys, for opening my eyes to this reality and the memories we made together! I am very grateful to the people in the aforementioned "bubble" and I would like to start the list by thanking my dear friends **Dima** and **Lena** for their support and understanding. You guys are the best and Bilge is lucky to have you in her life: future professor uncle who has answers and future wholesome artist aunt who does not need answers. Dear **Nigar** and **Musa**, you are the "go to" people for me when I need people who can get my jokes with a bright smile. Thanks a ton for being who you are. Deep thanks to **Kaspar** for being my personal advisor on local matters, for long walks and being such a perfect eestlane. I am grateful to **Sälly** for reminding me the value and coolness level of my academic work every time we meet. My dear gang of khinkalis: **Frozan**, **Tsotne**, **Anita**, **Sophie**, **Grete** and **Tair** - thanks for making Tartu a better place for me. **Volkan**, **Baran** and **Huri**, my globally approved friends, I consider myself extra-lucky for knowing and sharing a significant part of my life with you. Last but not least, thanks to my oldest friends **Jamal**, **Bahadur**, **Fezael**, **Faiq** and **Elgiz**. You are adding three additional levels of confidence to my life.

To my wonderful parents, **Terane** and **Ramazan**, I compared luck to a fair coin at the start of this thesis. In that metaphor, you two were undoubtedly my first stroke of luck, and I'm truly grateful to call you my parents. As the ones who brought me into this world, the success of this thesis is a testament to your achievement as well! I understand how tough it was to raise us and guide us to where we are now, and it seems all your efforts have paid off beautifully. It's truly hard to express how much I owe you, particularly parents as amazing as you. I hope my achievements bring you joy and pride. I love you! **Sizi sevirəm!**

Having someone you can talk to without holding back, and knowing they won't take it the wrong way, is truly priceless. My brother, **Tural**, is the go-to person for me, always there with open-hearted support and love, no strings attached. Thanks for **always** being there for me!

To my beloved daughter, **Bilge**: Thanks for being the light of my eyes! Without you, life just wouldn't be as meaningful.

I've mentioned a few times just how lucky I feel. My amazing wife, **Sariyya**, you're the biggest reason I feel this way. Every time I see you, every time you take care of Bilge on sleepless nights so I can rest, every time you stay home when Bilge is sick, every time you cheer me up when things get tough, every time you gently tell me to slow down when I'm pushing too hard, every time you remind me we're not competing when I feel I'm not doing enough, every time you give up your time and comfort to support me on this journey I've chosen—each of these moments is a testament to my luck. Every single time… **Çox sağ ol, sevgilIM!**

SUMMARY IN ESTONIAN

Komplekstunnuste geneetiliste seoste tõlgendamine molekulaarsete kvantitatiivse tunnuse lookuste andmebaasi abil

Käesolev doktoritöö keskendub peamiselt teaduslike meetodite täiustamisele ja rakendamisele, et selgitada inimesegeneetika uuringute kaudu komplekshaiguste ja -tunnuste aluseks olevaid bioloogilisi mehhanisme, kasutades eelkõige ülegenoomseid seosuuringuid (ingl k *genome-wide association studies* (GWAS)) ja kvantitatiivse tunnuse lookuste (ingl k *quantitative trait locus* (QTL)) analüüsi.

Komplekshaiguste ja -tunnustega on enamasti seotud suur hulk geneetilisi variante, millest igaühel on eraldi väikesed mõjud, mis omakorda muudab nende seoste tõlgendamise keeruliseks. Käesolevas doktoritöös kasutatakse GWAS uuringute abil tuvastatud geneetiliste variantide tõlgendamiseks QTL-analüüsi, et seostada need geneetilised variandid geeniekspressiooni tasemetega. Seejärel kasutatkse kolokalisatsioonianalüüsi, et veenduda, kas geeniekspressiooni ja komplekstunnuseid mõjutavad samad põhjuslikud variandid. See strateegia põhineb eeldusel, et komplekstunnust mõjutav geneetiline variant reguleerib algselt geeniekspressiooni. Kuna geeniregulatsioon varieerub sõltuvalt erinevatest bioloogilistest kontekstidest, sealhulgas rakutüüpidest, kudedest või tingimustest, võib kolokalisatsioonianalüüsi tegemine vales kontekstis muuta haigust või tunnust põhjustava seose tuvastamise võimatuks.

Doktoritöö annab põhjaliku ülevaate QTL analüüsi erinevatest tahkudest. Kõigepealt süvenetakse molekulaarbioloogia põhialustesse ja selgitatakse erinevaid lähenemisviise ja tehnikaid, mida kasutatakse GWASi signaalide tõlgendamiseks. Seejärel rõhutatakse erinevusi GWASi ja QTL uuringute vahel ja kirjeldatakse kolokalisatsiooni ja täppiskaardistamise (ingl k *fine mapping*) meetodeid. Samuti antakse ülevaade peamistest väljakutsetest, mis nende meetodite kasutamisel tekivad. Lõpuks antakse põhjalik ülevaade QTL analüüsi protsessist, mis hõlmab kogu töövoogu ja selle kolme põhikomponenti: 1) molekulaarsete tunnuse kvantifitseerimine, 2) kvaliteedikontroll ja normaliseerimine ning 3) QTLide kaardistamine. Üksikasjalikult käsitletakse RNA sekveneerimise lugemite arvulisteks andmetabliteks muutmise protsessi koos normaliseerimise meetodite nüansside uurimisega.

Teadusandmete analüüsi töövoogude kaasaegse taristu arendamine on olnud *eQTL Catalogue* andmebaasi loomisel esmatähtis. Doktoritöös kirjeldatakse teadusandmete töövoogude arenguetappe, käsitletakse arvutustaristu ja töövoogude omadusi ning antakse ülevaade pilvandmetöötluse ja tarkvara konteinerite tehnoloogiate edusammudest, mis on oluliselt suurendanud andmeanalüüsi töövoogude tõhusust ja skaleeritavust. Samuti tutvustatakse ja analüüsitakse töövoogude arendamise praeguseid väljakutseid, pakkudes võimalikke strateegiaid nende takistuste leevendamiseks. Lisaks käsitletakse põhjalikult

eQTL Catlaogue andmebaasi töövoogude arendamise põhiaspekte, tõstes esile protsessi käigus tehtud olulisi otsuseid ja nende hilisemat mõju töövoogude omadustele. Samuti antakse ülevaade töövoogude arendusprotsessi käigus sageli vähem tähelepanu saanud väljakutsetest ja õppetundidest. Viimaks antakse ülevaade QTLide visualiseerimise käigus tekkinud tehnilistest väljakutsetest ja nendele välja pakutud lahendustest.

Töö viimane peatükk annab ülevaate *eQTL Catalogue*'i tähtsusest teadusringkondade jaoks olulise ressursina, keskendudes selle erinevatele rakendustele paljudes teadusuuringutes ja sellele, kuidas teised teadlased on doktoritöös loodud andmebaasi kasutanud. See rõhutab jätkusuutliku ja säilenõtke taristu olulisust, et hõlbustada töövoogude kiiret rakendamist uutele andmestikele. Lõputöö võtab kokku kriitiline arutelu tehtud töö oluliste aspektide, selle peamiste piirangute ja võimalike edasiste arengute üle.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Nurlan Kerimov
Date of birth:	July 10, 1989
Citizenship:	Azerbaijan
Current position:	Junior Research Fellow of Bioinformatics
Contact:	kerimov.nurlan@gmail.com

Education

2019–2023	Ph.D. Candidate, University of Tartu
2017–2019	MSc. Software Engineering, University of Tartu and Tallinn Technical University
2007–2012	BSc. Electronics and Communications Engineering, Yildiz Technical University

Employment

2022–…	Lead Data Scientist, Nightingale Health Plc.
2019–…	Junior Research Fellow of Bioinformatics, Institute of Computer Science, University of Tartu
2018–2019	Scientific Programmer, Institute of Computer Science, University of Tartu
2014–2017	Software Development Engineer, Siemens
2012–2014	Software Developer, Farmasina Tıbbi ve Kimyevi Ürünler San. Tic. Ltd. Şti.

Teaching

- Trainer and teaching material co-creator of the course “Scientific Pipeline Development with Nextflow”
- Teaching assistant for the Bioinformatics course at University of Tartu in 2022

Scientific work

Main fields of interest: Bioinformatics, Scientific pipeline development, Genomics, Metabolomics

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Nurlan Kerimov
Sünniaeg:	10. juuli 1989
Kodakondsus:	Aserbaidžaan
Ametikoht:	Bioinformaatika nooremteadur
E-post:	kerimov.nurlan@gmail.com

Haridus

2019–2023	Tartu Ülikool, loodus- ja täppisteaduste valdkond, informaatika, doktoriõpe
2017–2019	Tartu Ülikool ja Tallinna Tehnikaülikool, loodus- ja täppisteaduste valdkond, tarkvaratehnika, magistriõpe
2007–2012	Yildiz Tehnikaülikool, elektroonika ja kommunikatsioonitehnika insener, bakalaureuseõpe

Teenistuskäik

2022–…	Juhtiv andmeteadlane, Nightingale Health Plc.
2019–…	Bioinformaatika nooremteadur, arvutiteaduse instituut, Tartu Ülikool
2018–2019	Teadusprogrammeerija, arvutiteaduse instituut, Tartu Ülikool
2014–2017	Tarkvaraarenduse insener, Siemens
2012–2014	Tarkvaraarendaja, Farmasina Tıbbi ve Kimyevi Ürünler San. Tic. Ltd. Şti.

Õpetamine

- Kursuse “Teaduslike töövoogude arendamine *Nextflow* abil” koolitaja ja õppematerjali kaasautor
- Bioinformaatika aine assistent Tartu Ülikoolis, 2022

Teadustöö

Peamised uurimisvaldkonnad: Bioinformaatika, teaduslike töövoogude arendamine, genoomika, metaboloomika

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.