# Optimisation of phasing: towards improved haplotype-based genetic investigations

by

Ziad Al Bkhetan

ORCID: 0000-0002-4032-5331

A thesis submitted in total fulfillment for the
degree of Doctor of Philosophy

in the
Melbourne School of Engineering
School of Computing and Information System
**THE UNIVERSITY OF MELBOURNE**

27 September 2020

THE UNIVERSITY OF MELBOURNE

# *Abstract*

Melbourne School of Engineering
School of Computing and Information System

Doctor of Philosophy

by Ziad Al Bkhetan
ORCID: 0000-0002-4032-5331

Haplotype or phase information significantly adds to the ability to resolve genetic problems and is important to elucidate and interpret certain genetic basis underlying diseases or traits. The common approach to derive phase information is through computational haplotype phasing or estimation methods. Current developments in phasing have paved the way to widely use haplotype information in population genetic investigations.

This PhD thesis explores various ways to utilise haplotype information effectively to conduct precise haplotype-based genetic investigations. It provides evidence of the important role of haplotypes to detect significant genetic associations with phenotypes that can be missed otherwise. In particular, it provides a comprehensive evaluation of state of art phasing approaches as well as different haplotype block determination methods (sliding window and through linkage disequilibrium). A rigorous analysis is conducted to improve phasing accuracy through a consensus haplotype estimator across datasets with different characteristics. Furthermore, phasing optimisation was utilised to develop a new approach to carry out haplotype-based Expression Quantitative Trait Loci (eQTL) analysis. The approach is assessed against genotype-based eQTL methods (both single and combinations of SNPs).

The main contributions of this PhD study are:

1. Novel evaluations and comparisons for haplotype phasing considering the accuracy at block scale that is the most popular way to use phase information in genetic studies.

2. An improvement of phasing accuracy reaching 10% when using the proposed consensus approach.

3. The consensus approach leads to the highest accuracy genotype imputation performed via the well-known tools Minimac3, pbwt and Beagle5.

4. An approach for haplotype-based eQTL analysis, that is demonstrated to outperform standard eQTL methods when the causal genetic architecture involves multiple variations.

Finally, the work in this PhD thesis highlights the fundamental role of haplotype information in genetic problems and provides guidance for other researchers interested in performing haplotype related investigations. Two tools (consHap and eQTLHap) are also released publicly with this PhD to support other research studies.

*To my family..*

# Declaration of Authorship

I, Ziad Al Bkhetan, declare that this thesis titled, 'Optimisation of phasing: towards improved haplotype-based genetic investigations' and the work presented in it are my own. I confirm that:

- The thesis comprises only my original work towards the degree of Doctor of Philosophy except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices as approved by the Research Higher Degrees Committee.

Signed: _____

Date: _____

# Preface

- This thesis includes the following publications that I was the primary author of.
  - The content of chapter 3 appears in this publication:
    Ziad Al Bkhetan, Justin Zobel, Adam Kowalczyk, Karin Verspoor, Benjamin Goudey. "Exploring effective approaches for haplotype block phasing." BMC Bioinformatics 20.1 (2019): 540. doi: 10.1186/s12859-019-3095-8.
  - The content of chapter 4 appears in a BioRxiv pre-print:
    Ziad Al Bkhetan, Gursharan Chana, Kotagiri Ramamohanarao, Karin Verspoor, Benjamin Goudey. "Evaluation of consensus strategies for haplotype phasing.". doi: 10.1101/2020.07.13.175786. It has also been published in Briefings in Bioinformatics - Oxford Academic (21 Sep 2020). doi: 10.1093/bib/bbaa280.
  - The content of chapter 5 appears in a manuscript under review in Briefings in Bioinformatics - Oxford Academic Journal (22 December 2020):
    Ziad Al Bkhetan, Gursharan Chana, Cheng Soon Ong, Benjamin Goudey, Ramamohanara Kotagiri. "eQTLHap: a tool for a comprehensive eQTL scan considering haplotypic and genotypic effects.". It also appears in a BioRxiv pre-print: doi 10.1101/2020.07.23.206391.
- Coeliac disease dataset used in Chapter 3 (EGA accession: EGAS00000000057) is generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.
- The Haplotype Reference Consortium (HRC) dataset used in Chapter 4 (reference: EGAD00001002729) is used in a form agreed by The University of Melbourne with Wellcome Trust Sanger Institute.

- The Genotype-Tissue Expression (GTEx) Project dataset used in Chapter 5 is used in a form agreed by The University of Melbourne with National Institute of Health (NIH). GTEx project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this thesis were obtained from dbGaP accession number phs000424.v8.p2 and from the GTEx Portal on 20-04-2020.
- The datasets from the 1000 Genomes project (used in Chapter 4), HAPMAP III (used in Chapters 3 and 4), and GEUVADIS (used in Chapter 5) are publicly available.
- This PhD project was supported by Melbourne Research Scholarship (ref: 103500) provided by the University of Melbourne.
- This PhD project was also supported by a top-up scholarship from Data61 started on 1$^{st}$ of July 2019 til the end of my PhD.
- This PhD project was conducted using computational resources supported by Melbourne Research Cloud (MRC).

# *Acknowledgements*

Writing this thesis would have proven a more challenging endavour without the support I have received from my supervisors and collaborators. I am very grateful to Gursharan Chana for his extraordinary support during various critical situations in the course of my PhD as well as in several other occasions. Words will not give him enough credit. I am thankful to Benjamin Goudey for helping me shape the direction of my PhD as well as for the valuable discussions and meetings we had. I would also like to express my sincere gratitude to both James Bailey and Rao Kotagiri for providing precious support in a critical situation in my PhD. Many thanks to Cheng Soon Ong for the invaluable discussions we had and for supporting my top-up scholarship application.

Many thanks to everyone who supported me including my family, friends and colleagues.

I gratefully acknowledge the HapMap III project, the 1000 genome project, the Haplotype Reference Consortium project, GEUVADIS project, and the Genotype-Tissue Expression (GTEx) Project including all unknown data donors.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **DNA** | **D**eoxyribo**N**ucleic **A**cid. |
| **SNP** | **S**ingle-**N**ucleotide **P**olymorphism. |
| **bp** | **B**ase **P**air. |
| **GWAS** | **G**enome-**W**ide **A**ssociation **S**tudies. |
| **eQTL** | **E**xpression **Q**uantitative **T**rait **L**oci. |
| **ASE** | **A**llele **S**pecific **E**xpression. |
| **LD** | **L**inkage **D**isequilibrium. |
| **HAPI-UR** | **HAP**lotype **I**nference for **U**n**R**elated samples. |
| **SHAPEIT** | **S**egmented **HAP**lotype **E**stimation and **I**mputation **T**ool. |
| **pbwt** | **P**ositional **B**urrows-**W**heeler **T**ransform. |
| **A** | **A**denine. |
| **G** | **G**uanine. |
| **C** | **C**ytosine. |
| **T** | **T**hymine. |
| **mRNA** | **M**essenger **R**ibo**N**ucleic **A**cid. |
| **U** | **U**racil. |
| **DFG** | **D**ifferential **G**ene **E**xpression. |
| **cDNA** | **c**omplementary **DNA**. |
| **NGS** | **N**ext-**G**eneration **S**equencing. |
| **T2D** | **T**ype **2** **D**iabetes. |
| **OMIM** | **O**nline **M**endelian **I**nheritance in **M**an. |
| **GAD** | **G**enetic **A**ssociation **D**atabase. |
| **IBD** | **I**dentity **B**y **D**escent. |
| **PCA** | **P**rinciple **C**omponent **A**nalysis. |
| **MAF** | **M**inor **A**llele **F**requency. |

| | |
|---|---|
| **HWE** | **H**ardy-**W**einberg **E**quilibrium. |
| **FPKM** | **F**ragments **P**er **K**ilobase of exon per **M**illion reads mapped. |
| **PEER** | **P**robabilistic **E**stimation of **E**xpression **R**esiduals. |
| **MHC** | **M**ajor **H**isto**C**ompatibility. |
| **MDD** | **M**ajor **D**epressive **D**isorder. |
| **AML** | **A**cute **M**yeloid **L**eukemia. |
| **AD** | **A**lzheimer's **D**isease. |
| **ASD** | **A**utism **S**pectrum **D**isorder. |
| **EGA** | **E**uropean **G**enome-phenome **A**rchive. |
| **EM** | **E**xpectation – **M**aximization. |
| **HMM** | **H**idden **M**arkov **M**odel. |
| **HRC** | **H**aplotype **R**eference **C**onsortium. |
| **SE** | **S**witch **E**rror. |
| **IHP** | **I**ncorrect **H**aplotype **P**ercentage. |
| **IGP** | **I**ncorrect **G**enotype **P**ercentage. |
| **ME** | **M**issing **E**rror. |
| **ANOVA** | **AN**alysis **O**f **VA**riance. |
| **htSNP** | **H**aplotype **T**agging **SNP**. |
| **GWHAS** | **G**enome-**W**ide **H**aplotype **A**ssociation **S**tudies. |
| **CEPH** | **C**entre d'**E**tude du **P**olymorphism **H**umain families reference panel. |
| **CEU** | **U**tah Residents from the **CE**PH collection with Northern and Western European Ancestry. |
| **IHBP** | **I**ncorrect **H**aplotype **B**lock **P**ercentage. |
| **IQR** | **I**nter**Q**uartile **R**ange. |
| **LCPR** | **L**ength of **C**orrectly **P**hased **R**uns. |
| **LDB** | IHBP calculated for **B**locks determined according to the **LD**. |
| **SWB** | IHBP calculated for **B**locks determined by **S**liding **W**indow. |
| **LDE** | **L**inkage **D**isequilibrium **E**rror. |
| **ASW** | **A**frican ancestry in **S**outh**W**est USA. |
| **CHB** | Han **CH**inese in **B**eijing, China. |
| **CHD** | **CH**inese in Metropolitan **D**enver, Colorado. |
| **GIH** | **G**ujarati **I**ndians in **H**ouston, Texas. |
| **JPT** | **J**a**P**anese in **T**okyo, Japan. |
| **LWK** | **L**uhya in **W**ebuye, **K**enya. |

| | |
|---|---|
| **MXL** | **M**e**X**ican ancestry in **L**os Angeles, California. |
| **MKK** | **M**aasai in **K**inyawa, **K**enya. |
| **TSI** | **T**oscani in **I**talia. |
| **YRI** | **Y**o**R**uba in **I**badan, Nigeria. |
| **GAM** | **G**eneralized **A**dditive **M**odel. |
| **CNV** | **C**opy **N**umber **V**ariations. |
| **TPR** | **T**rue**P**ositive **R**ate. |
| **BOW** | **B**ag **O**f **W**ords. |
| **FDR** | **F**alse **D**iscovery **R**ate. |
| **LOF** | **L**oss-**O**f-**F**unction. |
| **GTEx** | **G**enotype-**T**issue **Ex**pression. |
| **CI** | **C**onfidence **I**nterval. |
| **BH** | **B**enjamini-**H**ochberg. |
| **TSS** | **T**ranscriptional **S**tart **S**ite. |
| **PC** | **P**rincipal **C**omponents. |
| **B-Hap** | **B**lock's **H**aplotype. |
| **B-Gen** | **B**lock's **G**enotype. |
| **AIC** | **A**kaike **I**nformation **C**riterion. |
| **BH-qval** | **B**enjamini-**H**ochberg corrected pvalue. |
| **MTC** | **M**ultiple **T**est **C**orrection. |

# Chapter 1

# Introduction

## 1.1 Motivation

THREE major factors influence human traits including developing a disease: the environment they live in, the lifestyle they practise and the genome they inherit from their parents. This PhD project focuses only on the genomic component by exploring different genomic encoding (compared to the typical one) and investigates its association with traits through gene expression.

The genome within the cells of diploid organisms, such as humans, contains a set of chromosome pairs. Within a human cell, this set consists of 23 pairs of chromosomes where each copy of each pair is inherited from a single parent (Makałowski, 2001). Each copy of a chromosome pair is double-stranded series of four nucleotides termed *deoxyribonucleic acid (DNA)*. The instances of these nucleotides within each copy are termed *alleles*. The alleles allocated on the same copy are termed *haplotype*, while the combination of alleles within both copies represents the *genotype* of the organism.

Some genetic studies investigate variations in DNA sequences across individuals to find genomic loci associated with a specific phenotype (disease or trait). In the majority of these studies, the genotypic representation of the genome is considered for genomic variations and most conspicuously variations at a single base-pair termed *Single Nucleotide Polymorphisms (SNP)*. In comparison to such studies, there are very few investigations utilising the haplotypic representation of the genome, that is, considering the paternal

and maternal copies of each chromosome pair simultaneously. The scarcity of haplotype-based analyses can be related to many reasons, including the availability of haplotype data as well as the increase in complexity of the problem when dealing with the human genome at the haplotype level. Experimentally obtained haplotype information is not readily available by current genome sequencing technologies with an exception to the cases where the variations are allocated on the same sequenced read (Browning and Browning, 2011; Choi et al., 2018). However, these reads are short in general (up to 150 base pair (bp)) making haplotype assembly from several reads a difficult task (Snyder et al., 2015). Therefore, the main resource of haplotype information is computational methods called *haplotype phasing* or *estimation* methods. Using computationally phased haplotypes raises concerns about the impact of errors within the phased haplotypes on downstream analysis. Recent developments in haplotype phasing have led to high accuracy and scalability tools that are able to estimate haplotypes of large populations with very low error rate. Such improvements encourage the use of haplotype or phase information more widely in genetic studies.

An example of the studies mentioned above is *Genome-Wide Association Studies (GWAS)* that investigate genetic variations, most prominently SNPs, to reveal the genetic basis of a disease (Tam et al., 2019; Visscher et al., 2017). A large number of genetic variations have been reported to be associated with certain diseases (Buniello et al., 2019), with some demonstrated to be protective, while others conferring risk. Genetic variations can be associated with diseases in several ways such as introducing changes to the functionality of some associated genes (E.g. truncating variations can cause non-functional proteins) or at the level of their expression. The association between genetic variations and gene expression levels is assessed through a specific analysis termed *Expression Quantitative Trait Loci (eQTL)*. eQTL analysis aims at finding genetics variations that up/down-regulate the expression of particular genes. Such associations can provide a biological interpretation of how genetic variations influence a particular disease. It has been reported that a large percentage of genetic variations revealed by GWAS influence the expression of some genes (Jaffe et al., 2018; Nicolae et al., 2010). Due to its significance, eQTL analysis has been applied widely in genetic studies. Different approaches and statistical methods have been suggested for this purpose of considering single SNPs (Shabalin, 2012), epistasis (interaction of SNPs) (Hemani et al., 2014), and haplotypes

(Brown et al., 2017; Corradin et al., 2014; Garnier et al., 2013; Ying et al., 2018). Different genotype and gene expression datasets have been also made available to support developments of such approaches (GEUVADIS (Lappalainen et al., 2013) and GTEx (Consortium et al., 2015, 2017)).

The use of haplotype information for eQTL analysis has also been limited similarly to other DNA-based investigations. There are some known cases (termed *Compound Heterozygosity*) demonstrating that having specific mutations on the same copy of a particular gene can lead to a genetic disease (Tewhey et al., 2011; Zhong et al., 2017). Furthermore, the allocation of alleles on each haplotype copy can have a different impact on the gene expression of both homologous gene copies. For example, one copy is over-expressed compared to the other homologous copy (Knight, 2004; Tewhey et al., 2011). The latter case is investigated through *Allele Specific Expression (ASE)*. These examples demonstrate the significant impact of phase information in genetic studies that can lead to revealing associations, otherwise ignored by typical analysis. Since computationally estimated haplotype information is imperfect and has not been used widely in genetic investigations, there is an essential need to assess the limitations and strengths of computational haplotype phasing as well as to explore optimization of approaches to glean such information so that it may be utilised accurately and efficiently in genetic analyses, such as eQTL.

## 1.2 Knowledge gap

Using computationally phased haplotypes in genetic investigations raises concerns about the impact of errors within the phased haplotypes on the results of the downstream analyses, including eQTL analysis. Although recent phasing methods have been reported to be accurate, there is always a need to improve accuracy in order to conduct more authentic association analysis. This need becomes critical when dealing with eQTL analysis due to the small sample size of the available datasets (genotype and gene expression data for the same individuals) as the accuracy of phasing is dramatically reduced with small population size (Browning and Browning, 2011; Loh et al., 2016a; O'Connell et al., 2016).

The extra alleles' positioning details provided by phase information link different alleles within multiple loci on the same copy of a chromosome pair, therefore, there is no difference between the genotypic or haplotypic representation of a single SNP. The main advantage of haplotype information is when considering multiple variations. Partitioning the genome into blocks for analysis is a challenging task as the accuracy of phased haplotypes within the blocks is influenced by the way the genome is divided. Additionally, adding extraneous variations or eliminating important ones can affect the precision of the analysis. To date, there is no evaluation considering the accuracy of phasing at block scale. With the absence of such information, it is difficult to determine the optimal genome partitioning approach.

In the context of eQTL analysis, it is important to explore different approaches to conduct haplotype-based eQTL analysis as such approaches are not widely used. In addition, comparing findings of phase-based methods to the ones obtained by other approaches enriches our understanding of the genetic basis of up/downregulation of gene expression that can be linked to disease aetiology.

Finally, evaluation of the impact of phasing errors on eQTL analysis has not been reported. Such evaluations are crucial to avoid overconfidence in results of haplotype-based eQTL analysis.

## 1.3 Aims and hypothesis

We hypothesise that phasing accuracy can be improved by combining phased haplotypes obtained from different tools or multiple iterations of one tool. The improvements in accuracy facilitate the inclusion of phase information in association analysis. Including such information has the potential to elucidate novel relationships between genetic variations and diseases through the up- or down-regulate the expression of associated genes.

This study investigates the utility and limitations of statistical phasing in the context of genetic association; it explores the importance of including phase information to detect novel associations with a disease (through gene expression regulation) that cannot be captured via usual single SNP analysis, and it provides models and tools arising from this work to the genomics community.

The project addresses the following research questions:

1. What are the limitations and the expected phasing error within the haplotype blocks determined either via sliding window or based on Linkage Disequilibrium (LD) when phased by computational haplotype estimation tools?
2. How can we maximise phasing accuracy by aggregating multiple phased haplotypes into a consensus estimator applied to datasets with different characteristics?
3. Does including haplotype information in Expression Quantitative Trait Loci (eQTL) analysis lead to the detection of significant associations between genetic variations and gene expression that cannot be detected when analysing SNPs individually?

These aims are achieved by applying standard and novel evaluations of known and novel statistical phasing methods considering the accuracy of phasing at block scale (that is the most common way to use haplotype information in genetic problems). Inspired by findings from these evaluations, "optimal" phasing is carried out on genotype data to estimate individuals' haplotypes. The resulting haplotypes are passed as inputs into statistical methods to find haplotype blocks associated with the regulation of genes' expression.

## 1.4 Project significance

While this study is conducted fundamentally for a specific purpose that is the application of haplotypes for eQTL investigation, the reported findings and observations can be extended to other applications of haplotypes such as population genetic structure, genome-wide association studies and genotype imputation. We report novel evaluations and experimentation that can guide and support other research regarding computationally phased haplotypes. We freely provide computational tools to perform specific genetic tasks that can save researchers' time and effort.

This project focuses on optimizing phasing approaches to improve their accuracy and thereby increase the utility of phasing in genetic association studies, with a focus on the impact related to gene expression. As dealing with haplotypes for genetic problems is not common, this work and other related works highlight the benefits of such information

as well as motivate more frequent usage of it. The investigation of the proposed research questions leads to the following:

1. Finding the best scenario of combining haplotype phasing tools (individually or aggregated in a consensus estimator) and haplotype block determination methods in association analysis.
2. Achieving higher phasing accuracy that can lead to more accurate association analysis and obtain stable replicable results.
3. Providing evidence about the benefit of using haplotype information to detect novel associations between genetic variations and gene expression. Positive results will encourage to apply haplotype-based analysis to complement genotype-based analysis.

## 1.5 Thesis overview and contribution

The content of this thesis is structured in 6 chapters as follows:

1. **Chapter 1 - Introduction.**
   This chapter briefly introduces the investigated problem, the knowledge gaps, thesis's structure, aims and contributions.
2. **Chapter 2 - Background.**
   This chapter describes fundamental biological information and concepts needed to understand the biological perspective of the investigated problem as well as the source and the nature of the data used in the analysis. It sheds light on the knowledge gaps and discusses them deeply to highlight the contribution of the study and its significance. It covers statistics, techniques, and methods used within genetics that are related directly to the work done in this PhD. It also describes state of the art developments related to the problem and the motivation of the applied experiments and tests.
3. **Chapter 3 - Exploring effective approaches for haplotype block phasing.**
   This chapter introduces a majority voting consensus haplotype estimator (*termed consHap*) constructed from the outputs of different haplotype phasing tools. It

also provides a comprehensive evaluation of the proposed consensus estimator in addition to the state-of-art haplotype phasing tools (EAGLE2 (Loh et al., 2016a), HAPI-UR (Williams et al., 2012), SHAPEIT2 (Delaneau et al., 2012), BEAGLE (Browning and Browning, 2007b), IMPUTE (Howie et al., 2009), MaCH (Li et al., 2010), and fastPHASE (Scheet and Stephens, 2006)) with respect to phasing accuracy at block scale. The evaluation also considers two approaches to determine haplotype block boundaries (using sliding window and linkage disequilibrium) and suggests a novel metric to assess the accuracy within these blocks. The main contribution of this chapter is:

(a) A comprehensive evaluation of accuracy of phasing tools in the context of association analysis.

(b) A consensus estimator of SHAPEIT2, EAGLE2 and BEAGLE improves the accuracy by up to 10% compared to the best individual tool.

(c) Determining haplotype blocks based on linkage disequilibrium leads to more accurate blocks than a sliding window, but it provides a less comprehensive scan.

4. **Chapter 4 - Evaluation of consensus strategies for haplotype phasing.** This chapter investigates the consensus approach in more depth considering two different structures to construct the consensus (based on the output of different phasing tools, and the output of multiple application of one tool). The analysis is this chapter makes use of four haplotype phasing tools (SHAPEIT2, EAGLE2, SHAPEIT3 (O'Connell et al., 2016) and HAPI-UR) shown to be more accurate and efficient compared to other tools. It investigates the optimal structure to construct the consensus approach with respect to datasets with different characteristics, including population size and SNP density. It also provides an evaluation of the consensus approach according to different factors reported to have a high impact on phasing accuracy. Moreover, the impact of phasing on the downstream genotype imputation has been assessed and reported with respect to three genotype imputation tools (Minimac3 (McCarthy et al., 2016) that is used in Michigan imputation server, pbwt (Durbin, 2014) that is used in Sanger imputation server and Beagle5 (Browning et al., 2018)). The contribution of this chapter is:

(a) Exploring improvement of phasing accuracy through constructing consensus haplotype estimator based on majority voting of multiple phasing tools and multiple iterations of a non-deterministic tool. This evaluation considers datasets with different characteristics

(b) Assessment of impact of haplotype phasing on downstream genotype imputation conducted through BEAGLE5 (Browning et al., 2018), PBWT (Durbin, 2014) and Minimac3 (McCarthy et al., 2016).

(c) A consensus of multiple tools is almost always more accurate than a consensus of multiple iterations of a non-deterministic tool.

(d) The consensus approach achieves more accurate haplotype phasing and downstream genotype imputation in all experiments.

(e) *consHap*, a tool to construct the consensus haplotype estimator, was released freely.

5. **Chapter 5 - eQTLHap: a tool for a comprehensive eQTL scan considering haplotypic and genotypic effects.**
   In this chapter we exploit findings and observations from previous chapters to conduct haplotype-based eQTL analysis. Evaluations and analysis reported in both chapters 2 and 3 provide more details on using phased haplotypes blocks efficiently in association assessment. We address the impact of phasing errors on downstream eQTL analysis in order to demonstrate the validity of eQTL findings based on computational phasing. Supported by the previous chapters, we use a consensus approach for phasing as well as LD-based haplotype block determination. Blocks were assessed statistically to find significant associations with gene expression through a tool we developed (termed *eQTLHap*). The statistical assessment is extended to investigate the association with single SNP (typical eQTL analysis) as well as association with block's genotype. Results of the three approaches are compared using simulated data, then approaches are applied to real datasets of genotypes and gene expression (GEUVADIS and GTEx). The contribution of this chapter is:

   (a) We propose a new approach for haplotype-based eQTL analysis, termed *consHap*, and made it freely available online.

(b) We demonstrate that our approach outperforms typical SNP-based method for certain genetic architecture underlying the variation of gene expression.

(c) We show that phasing errors have a small impact on the true positive rate obtained by our approach ($< 4\%$).

(d) The application of our approach to real datasets from (GEUVADIS and GTEx) shows that eQTLHap discovers eQTL associations that can not be captured by SNP-base methods. These associations have been reported in other studies or tissues when compared with results available through eQTL catalogue (`https://www.ebi.ac.uk/eqtl/`).

6. **Chapter 6 - Discussion.**

   This chapter presents the main findings and conclusions of my thesis and discusses them in the context of the field of genetic-based investigations. Suggestions for future directions are given that may potential further advance the outcomes of this PhD.

# Chapter 2

# Background

This chapter provides a fundamental biological and computational details required to understand the remaining part of the thesis. It describes some terms and definitions in population genetics that cover the biological aspect of the problem. In addition, it includes technical details about state-of-art tools and approaches that are employed in this study and related works. In brief, this chapter explains the structure of the human genome, genotype and gene expression data, haplotype phasing and its evaluation, haplotype block determination methods, and expression quantitative trait loci (eQTL).

## 2.1 Introduction to genetics

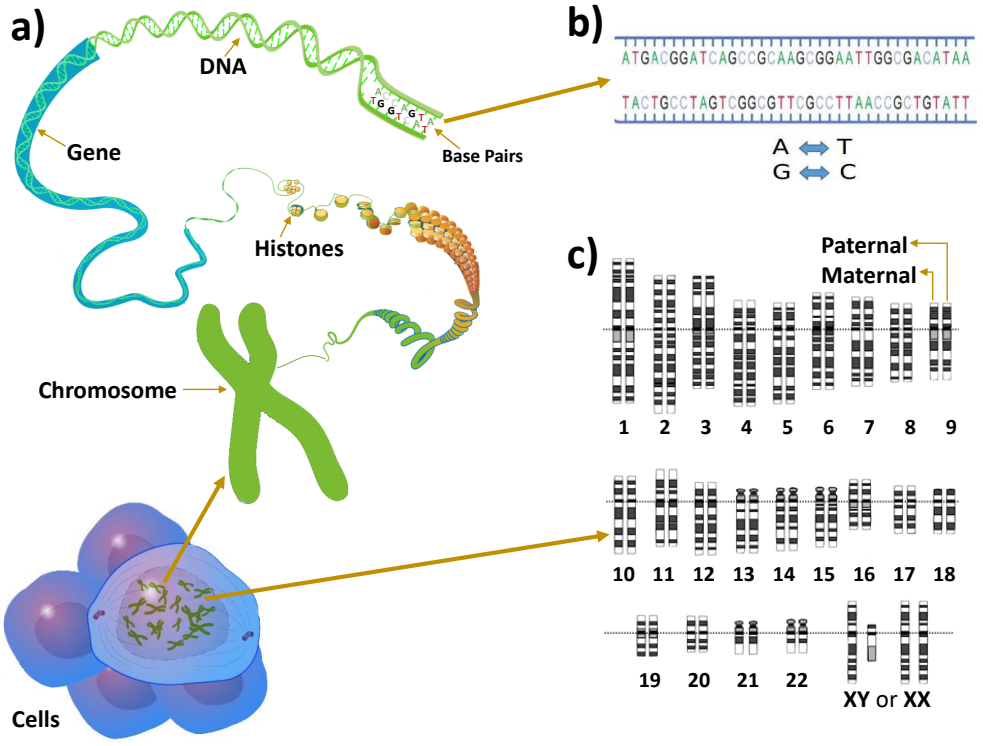
This section familiarises the reader with the biological aspect of this thesis and provides details about the nature of the data used in the analysis.

### 2.1.1 Human genome structure

The human genome is contained within the nucleus of all human eukaryotic cells and consists of 23 pairs of chromosomes (22 autosomes and 1 sex chromosome) (Makałowski, 2001) as illustrated in Figure 2.1. Each chromosome is composed of chaperone proteins, termed histones as well as helical double-stranded *deoxyribonucleic acid (DNA)* composed of nucleotides that are complementary to each other. The nucleotides are: *Adenine* (**A**), *Guanine* (**G**), *Cytosine* (**C**), and *Thymine* (**T**). The copies of each autosome pair are not identical but homological, that is, they contain the same set of

genes. Each copy of any chromosome pair is inherited from a single parent (paternal and maternal copies). Males inherit the sex chromosome X from their mother and the sex chromosome Y from their father, while females inherit two copies of chromosome X from their parents. Humans are considered diploid organisms for having two copies of each chromosome.

**Figure 2.1: Genome structure in the human nucleus.** a) Chromosomes within a cell and their structure. b) An example of the double-strands DNA sequence. c) Chromosome pairs in the human genome. This figure is adapted from two images credited to National Human Genome Research Institute `https://www.genome.gov/`.




### 2.1.2 From DNA to mRNA to protein

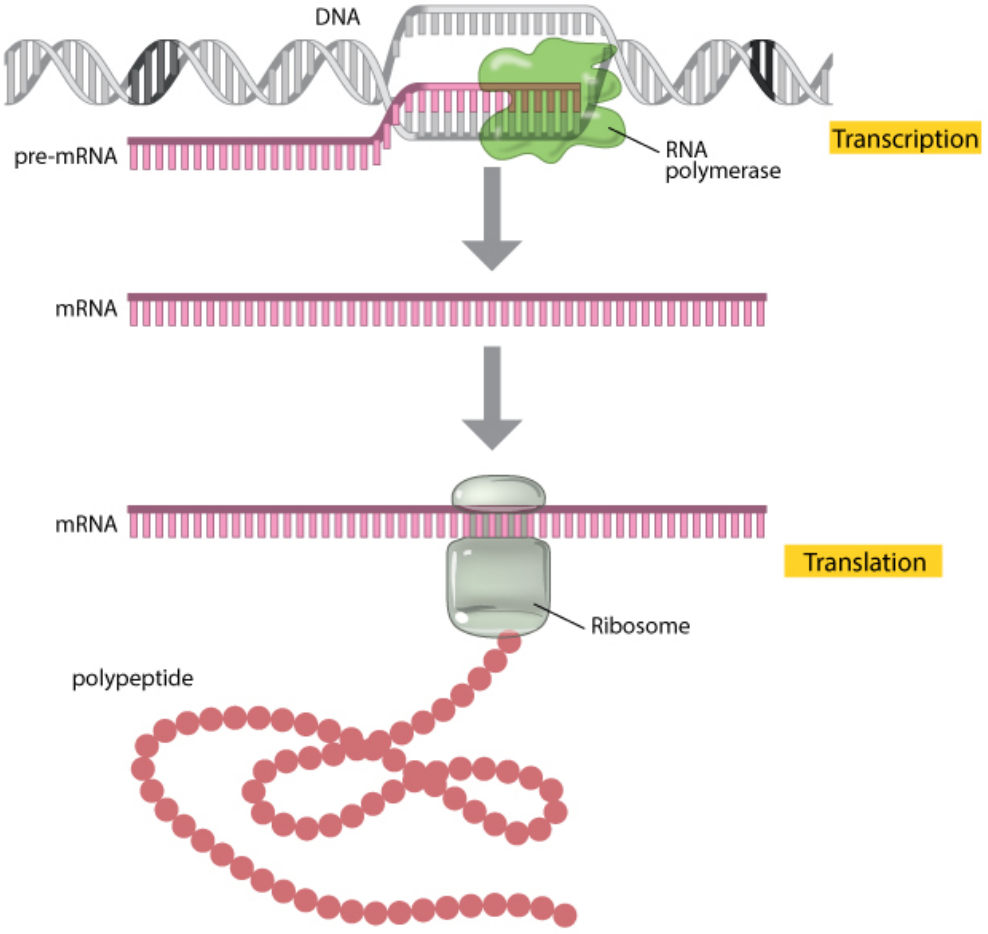
Proteins are considered the functioning elements in a cell, that is, they perform the majority of the cell's functions. Proteins are encoded by genes, that are specific regions of the DNA, through a process of two stages illustrated in Figure 2.2 (O'Connor et al., 2010):

1. **Transcription:** The DNA within a gene is transcribed into *messenger Ribonucleic Acid* (mRNA). Genes contain consecutive segments of DNA called *exons* and *introns*. Exons are the only segments that are transcribed. See Figure 2.5.

The formed mRNA strand is the same as the complementary strand of the transcribed DNA, but with a replacement of thymine (T) by *uracil* (U). Transcription is carried out by an enzyme called *RNA polymerase II*. RNA polymerase II binds to the DNA in specific regions, called regulatory regions, such as promoters and enhancers, then initiates the transcription process.

2. **Translation:** mRNA is translated into a protein. Each three RNA bases termed triplet codons are translated into an amino acid. There are 64 possible triplet codons (possible combinations of three molecules out of four: A, G, C, and U) that encode all 22 amino acid including 2 non-standard amino acid.

**Figure 2.2: Transcription and translation processes.** RNA polymerase enzyme binds to the DNA to initiate the transcription process. The DNA is used as a template by RNA polymerase to form mRNA molecule that is translated into protein by a complex molecule called ribosome. This figure is credited to Nature Education `https://www.nature.com/`.



### 2.1.3 Genomic variations

The human genome is approximately two meters long (Anthony T., 2008) and contains more than 3 billion nucleotides (Jackson et al., 2018) that encode 20-25,000 genes (Sealfon and Chu, 2011). It is estimated that approximately 0.1% of the genome varies between individuals (Consortium et al., 2012a; Jorde and Wooding, 2004). Most genetic studies only focus on these variations to understand their impact on different traits or diseases. Genomic variations can exist in several forms, depending on the size and the type of variation:

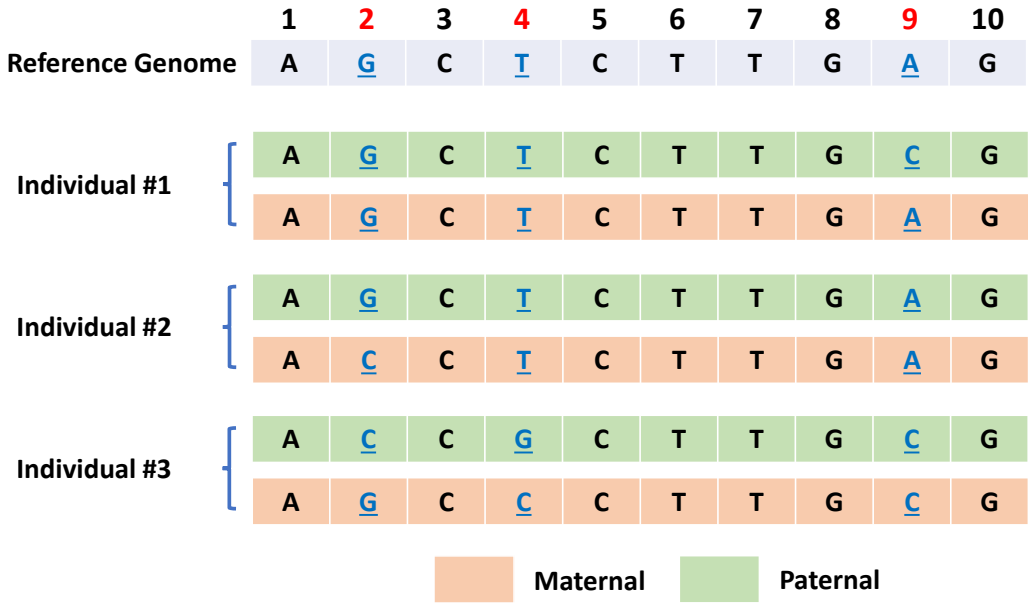
- **Structural variations** happen at a large scale (> 1k base pair (bp)) (Jackson et al., 2018), such as chromosome rearrangement, and copy number variations.
- **Insertion and deletion variations** (indels) happen at a smaller scale (< 1kbp).
- **Repeat variations** which are duplicated genomic regions within the genome.
- **Nucleotide substitutions or Single Nucleotide Polymorphisms** (SNPs) that occur at a single base scale. 90% of genomic variations are SNPs (Collins et al., 1998). A SNP is a variation in a single base-pair across a population. There are roughly 10 million SNPs in the human genome with an average of 1 SNP every 300 nucleotides (Jackson et al., 2018).

### 2.1.4 SNP's alleles and haplotypes

Where a SNP occurs on a chromosome, the SNP is described to have more than an *allele* at its locus. Each instance of this SNP on each copy of the chromosome pair is an allele, therefore, a SNP can have up to two different alleles within each individual, but it can be more than two when considering multiple individuals at the same genomic locus. The allele identical to the one in the reference genome (a sequence of common alleles across a population) within the same locus is called the *reference allele* otherwise it is *alternate allele*. The alleles inherited together from the same parent, or in other words located on the same copy of a chromosome pair, are called *haplotype* or *phase*. The combination of alleles reflective of genomic variations is make up the *genotype* of an organism. A SNP within an individual is termed either *homozygous* when both of its alleles are identical (regardless whether they are reference or alternate alleles), or *heterozygous* when the

alleles differ. When a SNP has only two possible alleles across the whole population it is termed *biallelic* SNP. Figure 2.3 illustrates an example of genomic data for a population that clarifies all mentioned definitions. This PhD study focuses only on biallelic SNPs and more precisely considering the haplotypes within these SNPs.

**Figure 2.3: An example of genomic sequences in a population.** Reference genome: a sequence of the common alleles (most frequent) within the specified population. Each individual (of 1, 2, and 3) has a pair of chromosomes. We assume that green copies are paternal (inherited from the father) while the orange is maternal (inherited from the mother). Each nucleotide on any of these copies is an allele such as "*G*" in the paternal copy of the individual #2 at the locus 8. When compared to the alleles of the reference sequence, the loci 2, 4 and 9 represent SNPs as there is a variation across the whole population. The loci 1, 3, 5, 6, 7, 8, and 10 are ignored in the majority of the genetic studies as they are identical across all individuals. The alleles "*CGC*" at the loci 2, 4, 9 or the complete allele sequence "*ACCGCTTGCG*" represent the paternal haplotype of the individual #3 depending on the considered loci. The combination of the alleles "*(AA)(CG)(CC)(CG)(CC)(TT)(TT)(GG)(CC)(GG)*" makes up the genotype of the same individual with respect to the loci (having in mind that the allocation of the alleles is ignored in the genotypic representation). The allele "*A*" at the locus 9 of the maternal haplotype of the individual #1 is an example of a reference allele, while the paternal allele "*C*" at the same locus is an alternate allele. The SNPs at locus 9 for both individuals 2 and 3 are homozygous while it is a heterozygous SNP at the same locus for the individual 1. SNPs at loci 2 (alleles C and G) and 9 (alleles A and C) are biallelic SNPs as they only have two alleles across all individuals, while it is a triallelic SNP at the locus 4 (alleles C, G and T).




### 2.1.5 Linkage disequilibrium (LD)

Neighbouring alleles are more likely to be inherited together, therefore, the correlation between them is greater than what would be expected by chance. This correlation or

linkage is termed *linkage disequilibrium* (LD). Consider two SNPs $S_1$ and $S_2$, with major and minor alleles {A, a} and {B, b}, respectively. If A and B are independent alleles (no correlation), the probability of observing the haplotype AB is equivalent to the product of observing each allele separately. The LD between the alleles A and B can be measured through the deviation of the probability of observing the haplotype AB from the probability of observing both alleles A and B as independent alleles (Balding et al., 2008). See Equation 2.1.

$$D_{AB} = P_{AB} - P_A \times P_B \tag{2.1}$$

The probability of observing two independent events together is: $P(E_1 \cap E_2) = P(E_1) \times P(E_2)$.

When $D_{AB} > 0$, the alleles A and B are observed together more than what is expected if they were independent. In other words, they are correlated. With respect to the genotype data of a population, the probabilities are calculated based on the frequencies within the population. $D$ is highly influenced by the frequency of the alleles (Balding et al., 2008). $r^2$ is another way to estimate LD, that is less sensitive to allele frequency (Balding et al., 2008). $r^2$ is equivalent to the squared Pearson correlation and can be defined as equation 2.2.

$$r^2 = \frac{D_{AB}^2}{P_A P_B P_a P_b} \tag{2.2}$$

D-prime ($D\prime$) (Lewontin, 1964) is another metric to calculate LD that is less influenced by allele frequencies than $D$. It can be calculated as follow (Balding et al., 2008):

$$|D'_{AB}| = \begin{cases} \frac{|D_{AB}|}{min(P_A P_b, P_a P_B)}, & \text{if } D_{AB} > 0 \\ \frac{|D_{AB}|}{min(P_A P_B, P_a P_b)}, & \text{if } D_{AB} < 0 \end{cases} \tag{2.3}$$

$|D'|$ takes a value between and including 0 and 1, where $|D'| = 1$ is an evidence for no recombination (termed *complete LD*), while a value near 0 can be an indication for a considerable recombination (Balding et al., 2008).

LD decreases when the genomic distance between the alleles increases (Browning and Browning, 2007b) as the chance of recombination or crossover events increase. It is important to note that with the absence of real haplotype information, that is common, LD is calculated based on computationally estimated haplotypes with respect to the observed genotype data.

### 2.1.6 Gene expression

The DNA contains the genetic instructions required for cells to function, however, the cells of the same individual (that contain the same copy of DNA, in other words, the same genetic instructions) can behave in a different manner. The reason for that is, proteins, rather than the DNA of a gene, are responsible for the functionalities of the cell. Measuring the functional level of protein present in a cell is very important to understand and assess the functionality of a cell or a tissue, yet it is a hard procedure. Alternatively, it is possible to measure mRNA level within a cell which gives an estimation of protein levels as mRNA is translated into protein (Sealfon and Chu, 2011). Across different cells, not all genes are expressed into mRNA, and when they do, it is not necessarily that they are expressed in a similar manner. With respect to a particular cell, genes are considered either active when they are highly expressed, or inactive when they are negligibly expressed (low expression).

A specific type of genetic studies, such as *differential gene expression* (DFG) studies (Carulli et al., 1998; Liang and Pardee, 2003; Robinson et al., 2010), investigate the difference of the gene expression across different instances to reveal a potential impact on a specific condition. Genes are termed differentially expressed when there is a statistically significant difference between their expression levels. Gene expression analysis can be applied to cells of different tissues of the same individual, healthy and tumour cells, or the same cell of different individuals (case/control).

### 2.1.7 Variant detection and gene expression profiling

Both variant detection and gene expression profiling can be carried out via microarray or sequencing technologies. Microarray technology was introduced in 1995 (Schena et al., 1995). Microarrays consist of a chip containing thousands of tiny spots each of them representing a set of probes of a known DNA fragment or gene that complement the targeted sample. There are several platforms for microarrays, all these platforms use a similar procedure to detect a SNP or to measure gene expression level. When detecting SNPs, several hybridised DNA probes (Single DNA strand that complements a short fragment that contains the targeted SNP) are used to capture the possible alleles of an

individual's specific SNPs. DNA is collected from cells of an individual and then denatured into two complementary strands of DNA. DNA strands are partitioned into small fragments then amplified using RT-PCR and labelled using fluorescent dyes. Samples are mixed and applied to the chip to allow sample's fragments to bind to specific DNA probe sets located on the microarray chip. The binding process is termed *hybridization*. The samples are washed from the chip, then spots are scanned via laser to determine their intensity, allowing the identification of SNPs.

Gene expression can be measured via microarray similarly to the procedure of SNP detection. mRNA is extracted from a tissue or cell line for both target sample and reference sample (such as case/control) when using a two-colour or two-channel array. The two mRNA samples are then converted into complementary DNA (cDNA) and labelled using fluorescent dyes. Following hybridisation of cDNA, chips are washed and the array is scanned and analysed via a laser to determine the gene expression level based on the intensity of fluorescence of both the target and reference samples. While the two-channel array allows comparing two samples directly from the same microarray, it is also possible to measure both samples separately using a one channel micro-array using a similar procedure.

The most recent technology used for SNP detection and gene expression profiling is genome sequencing. Sanger sequencing was the first method to sequence short reads of the genome. It was developed by the British biochemist Fred Sanger and his colleagues in 1977, for which he received the Nobel prize. Recently, *next-generation sequencing* (NGS) can do whole-genome sequencing and provide long sequenced reads (reaching thousands of bases (Buermans and Den Dunnen, 2014; McCombie et al., 2018)), with costs for such sequencing continuing to decrease with the increased adoption of the technology. Variants detection or gene expression profiling using sequencing technology require a reference sequence. DNA sequenced reads (after reverse transcription for RNA when gene expression is measured) are aligned to the reference genome based on the sequence similarity. Different alleles within a specific locus or the count of aligned reads to gene's exons can be identified for SNP detection and gene expression measuring, respectively. Genome assembly from the sequenced reads can be done if there is no reference sequence. There are some methods to detect SNPs from the sequenced reads directly without the need to a reference sequence.

Even though NGS is the recent technology in this field, microarrays are still used widely due to their reliable results obtained in the last two decades. Table 2.1 shows a comparison of some features of both technologies.

**Table 2.1: Comparison of microarray and genome sequencing technologies.**

| Feature | Micro-array | Sequencing |
|---|---|---|
| **Prior knowledge about the sequence** | Requires | Does not require |
| **Resolution** | Several bp to 100 bp | Single bp |
| **Cost** | Cheaper | More expensive |
| **Novel variants detection** | Not possible | Possible |
| **Structural variants detection** | Not possible | Possible |
| **Gene expression quantification** | Derived from intensities in micro-array image | Derived from read count |
| **Usability** | Proven track record | New to most researchers |

Although, the considerable development in this field, haplotype information is not readily available (Choi et al., 2018). Microarrays can not provide any information about the allele location on each copy of a chromosome pair. Genome sequencing can resolve this issue partially when multiple SNPs are allocated on the same sequenced read (Browning and Browning, 2011). However, the reads are relatively short when compared to the length of the whole chromosome. Haplotypes can not be determined when the SNPs are allocated on different reads. Recently, NGS can do whole genome phasing (called *phased sequencing* [1]). Despite the experimental methods, the main resource of haplotype information is computational approaches called haplotype estimation or phasing (will be explained in details later in this chapter in the section: 2.4 Haplotype phasing). Haplotype estimation methods phase the genotype data of a population and separate the paternal and maternal alleles (Browning and Browning, 2011).

We refer the reader to these reviews for more details related to both technologies including their history and the main milestones in their development (França et al., 2002; Heather and Chain, 2016; Heller, 2002; Mardis, 2008; Miller and Tang, 2009; Shendure et al., 2017).

[1] `https://sapac.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing/phased-sequencing.html`

### 2.1.8 The genetic basis of a phenotype

The phenotype of an organism refers to its observable characteristics such as having a disease or a specific trait. Generally, disease pathogenesis and other traits involve the influence of both environmental and genetic factors (Gohlke et al., 2009; Jirtle and Skinner, 2007). This convoluted relationship makes it hard in many instances to determine the pure genetic impact on a phenotype. The genetic basis underlying disease or trait is termed *heritability*, that is, the percentage of phenotypic variation that can be explained by the genetic variation. There are many genetic components of disease heritability (Jackson et al., 2018), such as:

1. Genomics variations including SNPs, insertion, deletion and structural variations as described in section 2.1.3 Genomic variations.
2. Chromosomal imbalances.
3. Epigenetics that includes DNA methylation and histone modifications.

There are many ways that genomic variations can be associated with a disease. Some variations impact the functionality of a protein by changing its structure (Studer et al., 2013) such as introducing a stop codon that truncates the protein (shortened protein) or substituting an amino acid by another one in a way that affects the functionality of the protein. It has been estimated that the number of functional genes across individuals can vary by up to 10% due to the impact of SNPs (Jackson et al., 2018). Other variations can impact the regulation of specific genes (Francesconi and Lehner, 2014; Tewhey et al., 2011). Usually, such variations are allocated on the regulatory regions.

Disorders are categorised into single-gene or complex disorders according to the complexity of their underlying genetic aetiology.

- **Single-gene disorders**: They are also termed *Mendelian* or *monogonic* disorders. These disorders are associated with a single SNP or multiple SNPs allocated on the same gene.
- **Complex disorders**: They are also termed *polygenic* disorders. In these disorders, the impact of a single variation is very small, however, the contribution

of multiple variations on other genes, in addition to a particular environmental context can increase the risk factor of disease. Examples for these disorders are Type 2 diabetes (T2D), heart disease and schizophrenia. There are some traits and behaviours that are considered complex as well such as height and aggression (Jackson et al., 2018).

The genetic contribution to disease aetiology has been investigated extensively in the last few decades. Numerous studies have revealed direct genetic-disease associations at both molecular scales DNA (i.e. genetic variations) (Buniello et al., 2018) and RNA (i.e. gene expression) (Emilsson et al., 2008; Schadt et al., 2005) or indirectly through the detection of variations influencing the regulation of particular genes that can be associated with disease (Albert and Kruglyak, 2015; Lonsdale et al., 2013). Considerable efforts have also been spent to integrate information about genetic associations with diseases in a centralised resource, that can be easily accessed by researchers. GWAS catalog [2] contains 3,314 publications of 59,145 unique SNP-trait associations (Buniello et al., 2018). Online Mendelian Inheritance in Man (OMIM) [3] has reported that 6,447 phenotypes have known molecular basis, that includes 5,390 single-gene phenotypes and 694 complex phenotypes (OMIM statistics were obtained in 21st June 2019). The Genetic Association Database (GAD) [4] provides information about the genetic basis of many complex diseases and disorders (Becker et al., 2004). Table 2.2 includes some example of well-known genetic associations with diseases or traits.

## 2.2 Quality control procedure and data preparation

When analysing genetic data such as genotypes or gene expression, some standard pre-checks should be applied as quality control to ensure the data is robust for the intended analysis. Several errors can occur when obtaining genetic data from experimental methods, with some being caused by technical reasons related to the used equipment, or biological reasons such as un-modelled biological differences between samples and a variation in the DNA sequence. Such errors were observed and investigated previously

[2] `https://www.ebi.ac.uk/gwas/home`
[3] `https://www.omim.org/`
[4] `https://geneticassociationdb.nih.gov/`

**Table 2.2: Examples of well-known genetic associations with diseases or traits.**

| Variation | Phenotype | Ref |
|---|---|---|
| rs7495174, rs6497268, and rs11855019 SNPs. | Blue/nonblue eye color. | (Duffy et al., 2007) |
| Variations on the BRCA1 or BRCA2 genes. | Increases the risk of developing breast cancer and ovarian cancer. | (Antoniou et al., 2003) |
| A deletion of three nucleotide on the gene CFTR. | Associated with Cystic fibrosis disorder. | (Kerem et al., 1989) |
| T allele of SNP rs7903146 on gene TCF7L2. | Strong risk factor for T2D. Associated with better response to specific T2D medication. | (Jackson et al., 2018) |
| rs4430796 SNP on HNF1B gene. | Protective for T2D. Associated with increased risk of prostate cancer. | (Jackson et al., 2018) |
| rs121434622 SNP on FMR1 gene. | Associated with Fragile X syndrome (monogenic). | (De Boulle et al., 1993) |
| rs28934907 SNP on MECP2 gene. | Associated with Rett syndrome (monogenic). | (Amir et al., 1999) |
| rs77543610 on FGFR2 gene. | Associated with Apert syndrome (monogenic). | (Wilkie et al., 1995) |

for genotype (Pompanon et al., 2005) and gene expression (Rocke and Durbin, 2001; Weng et al., 2006) data.

### 2.2.1 Pre-analysis checks

Listed below are a few important checks that should be applied to genotype or gene expression data.

1. **Gender inconsistency:**

   It is very important to verify the gender of all individuals in the dataset. An individual's gender can be predicted from genotype data, then compared to the gender reported for each individual from the data source.

2. **Sample relatedness:**

   Data should be investigated to verify whether there is any unknown relatedness among the individuals included in the dataset. Related individuals who have high similarity in their genotypes can be a confounder in association analysis. Such investigation usually relies on finding some long genomic region (called *identity by descent* (IBD) ) shared across some individuals. Relatedness can be checked for each pair of individuals within the dataset, by assessing the probability for two individual to share 0, 1, or 2 IBD alleles (Laurie et al., 2010; Turner et al., 2011).

3. **Population stratification:**

   This is a well-known issue when applying association analysis especially GWAS. It refers to the case where the individuals within a dataset can be clustered into different sub-populations due to substantial allele frequency differences. A visualisation based on principle component analysis (PCA) can show if there is population structure in the dataset. It is recommended to prune the SNPs of genotype dataset leaving low correlation SNPs before applying PCA (Turner et al., 2011). Eliminating highly correlated SNPs helps to avoid any potential bias in PCA analysis.

4. **Batch effect:**

   Thousands of samples are assayed together in batches when applying genotype calling or gene expression profiling. As the random sampling for each batch may vary, there is a potential impact on the results. The batch effect can be investigated through PCA similarly to population stratification (Turner et al., 2011).

### 2.2.2 Genotype data quality control

The quality of genotype data is one of the main concerns when conducting any genetic analysis. Even with a low error rate, there can be a large number of incorrectly genotyped SNPs. Such errors can impact any downstream analysis and increase the false positives in the results (Laurie et al., 2010; Wang, 2018). It has been reported that 16% of SNPs available in public databases and unconfirmed by other studies are caused by sequencing errors and low allele frequency (Pompanon et al., 2005). Furthermore, it has been estimated to have from 0.2% to 15% genotype error per locus (Bonin et al., 2004; Wang, 2018). With the increase genotypes sample size, the error rate is more likely to increase even when using next-generation sequencing technology (Wang, 2018). There is also a high error rate when SNPs are called from low-coverage NGS data (Nielsen et al., 2011; Wang, 2018). To avoid such issues, several approaches have been proposed for genotype data cleaning and preparation.

1. **Missing genotypes or call rate check:**

   It is recommended to eliminate any SNP if it is not genotyped in at least 95-99% of the available samples (Laurie et al., 2010; Turner et al., 2011). Different thresholds were proposed by different studies, however, the higher the threshold the more accurate the data.

2. **Mendelian errors:**

   In some cases and due to some genotyping errors, there can be allele inconsistency across individuals from the same family. For example, a child has an allele "A" in a specific locus, when both parents do not have the same allele at the same locus. Such cases can be detected when family data is accessible, and they are then preferentially eliminated.

3. **Minor allele frequency (MAF) filtration:**

   It has been reported that the chance of association detection with rare SNPs is extremely low, therefore, some studies exclude SNPs with ($MAF < 1\%$) (Turner et al., 2011). It is important to note that haplotype-based association analysis has been reported to be a powerful approach to detect associations with rare SNPs (Howard et al., 2017), therefore, this filtration can be relaxed when applying such analysis.

4. **Hardy-Weinberg equilibrium (HWE):**

   The deviation from Hardy-Weinberg equilibrium (HWE) is an indication of genotype errors, population stratification and disease susceptibility loci (Ryckman and Williams, 2008; Turner et al., 2011). This deviation can be calculated by comparing genotype frequencies to the expected frequencies under HWE that assumes random mating within a population and sample independence (Goudey, 2016). More details about applying HWE-quality control are reported in this study (Ryckman and Williams, 2008).

PLINK (Purcell et al., 2007) is a well known publicly available software that can help to conduct quality control procedures. It is available at this link `https://www.cog-genomics.org/plink`.

### 2.2.3 Gene expression normalisation

Gene expression profiling through sequencing approaches reports the total number of short reads that are aligned to each gene with respect to a reference genome. Therefore, there is a high chance that the total number will be large for long genes compared to short ones. With different experiments, the size of the RNA library used by the experimental method changes, therefore, the counts can vary accordingly. For example, it is

expected that gene expression levels obtained by applying an RNA library of 50,000,000 reads will be 5 times the results obtained by using a library of 10,000,000 reads. Accurate gene expression analysis should account for both factors by dividing the total count by the length of the gene and the size of the RNA library. This measure is termed FPKM (fragments per kilobase of exon per million reads mapped). However, this normalisation approach does not account for other resources of gene expression variation including biological, environmental, and technical factors such as batch effect, population stratification, sample history and environmental conditions. Such factors can influence gene expression profiling and impact downstream analysis such as eQTL (Balding et al., 2008; Stegle et al., 2012). While such factors can be included in the statistical models used as covariates to eliminate such impact, this scenario requires the factors to be determined ahead before. Unfortunately, it is not the common case and most of these factors are unknown.

Probabilistic estimation of expression residuals (PEER) is a normalisation method that accounts for such variation in gene expression (Stegle et al., 2012). PEER approach relies on an approximate inference to determine hidden expression determinants. These determinants can be used as covariates (in addition to other known ones) in the statistical assessment. Another way to use PEER is to apply any statistical assessment (such as eQTL) on the residual dataset instead of the original gene expression dataset as PEER's residual dataset is equivalent to the original gene expression dataset yet after removing the contribution of the determinants. PEER approach has should be applied on gene expression data after quality control and FPKM normalisation procedures.

Other normalisation approaches account for sample bias, library size and gene length (Abbas-Aghababazadeh et al., 2018). We refer the reader to this comparison paper that covers gene expression normalisation (Abbas-Aghababazadeh et al., 2018). There are also some methods to normalise gene expression data obtained by microarrays (Park et al., 2003).

## 2.3 The importance of haplotype/phase information

When investigating the genetic basis of a disease, most studies have ignored haplotype information and focused on the combination of alleles within each genomic locus (genotype). In doing so, the link between the alleles allocated on each copy of a chromosome pair is lost, and this missing information can limit the ability to fully understand the genetic contribution to disease. The allocation of the alleles on the homologous copies of each chromosome does not matter when dealing with homozygous SNPs as the alleles are identical. However, it is not the case when dealing with heterozygous SNPs whose alleles are different and can have a different impact according to their allocation on the chromosome copies.

### 2.3.1 Limitations of genotype-based analysis

Most genetic association studies are conducted at the genotype level, starting from the assessment of each SNP separately, followed by the incorporation of more sophisticated methods that assess the interaction between multiple SNPs that can influence a particular phenotype but that separately do not have the same impact. While these methods or approaches have the ability to successfully identify significant genetic associations with a disease where a single gene or locus or interaction between a few genes can impact on the phenotype, the majority of diseases are genetically more complex and are associated with multiple SNPs in multiple genes. SNP-based methods suffer from many computational and technical restrictions and limitations when trying to assess multiple SNPs together due to the huge number of tests needed to assess all possible combinations of SNPs (Carmelo et al., 2018; Niel et al., 2015). Two SNPs analysis is achievable at present through current computers, but the difficulty increases exponentially when trying to assess a greater number of SNPs. An exhaustive evaluation for over 500 billion SNP pairs in genome-wide was done before (Goudey et al., 2013). It is estimated that three-SNPs interactions analysis on 1.1 million SNP requires over 5.8 years (Goudey et al., 2015). Also, SNP-based analysis can not combine the information from different SNPs efficiently, and it is hard for such analysis to capture the association with low-frequency variants (Howard et al., 2017).

### 2.3.2 Advantages of using haplotype information

Advantages of using haplotype information in genetic studies are explained below:

- Previous studies have demonstrated that haplotype stretches are shared between individuals (Kong et al., 2008; Li and Stephens, 2003; Yang et al., 2008). The possible haplotypes of a typical gene are 10 to 15 with a few exceptions (such as major histocompatibility (MHC) genes in chromosome 6) (Yang et al., 2008). The genotype data representing a genomic region does not necessarily reflect the actual haplotypes within the region. Different genotype sequences can be formed by a small set of different haplotypes when haplotype pairs are mated randomly. In this study, we found that there exist 343 different genotype sequences within a region of 354 SNPs in chromosome 1 across 373 individual of the European population from 1000 genome project (their frequencies vary from 0.13% to 0.5%). After phasing this region, we found only 252 different haplotypes (28% reduction) within the same region (their frequencies vary from 0.13% to 5%. Using haplotype information can facilitate the analysis of multiple SNPs dramatically by adding allocation detail to genetic variations.

- Analysing haplotype blocks instead of SNPs reduces the number of statistical tests dramatically as the number of haplotype blocks is significantly less than the number of SNPs. The reduced number of statistical tests provides less conservative significance thresholds and more flexibility when correcting for multiple comparison (Yang et al., 2008; Ying et al., 2018).

- The analysis of haplotypes means dealing with the real DNA sequences rather than focusing on genomic variations without considering the alleles location. The knowledge of the actual alleles allocated on the same chromosome can be transferred into the knowledge of the translated amino acids when the alleles are allocated on a gene. Understanding the impact of the variations on the translated protein greatly helps to understand disease aetiology.

### 2.3.3 Applications of phased haplotypes

Several known cases show the importance of haplotype or phase information in genetic studies (Crawford and Nickerson, 2005; Tewhey et al., 2011; Tregouet and Garelle, 2007). Examples of these cases are explained below:

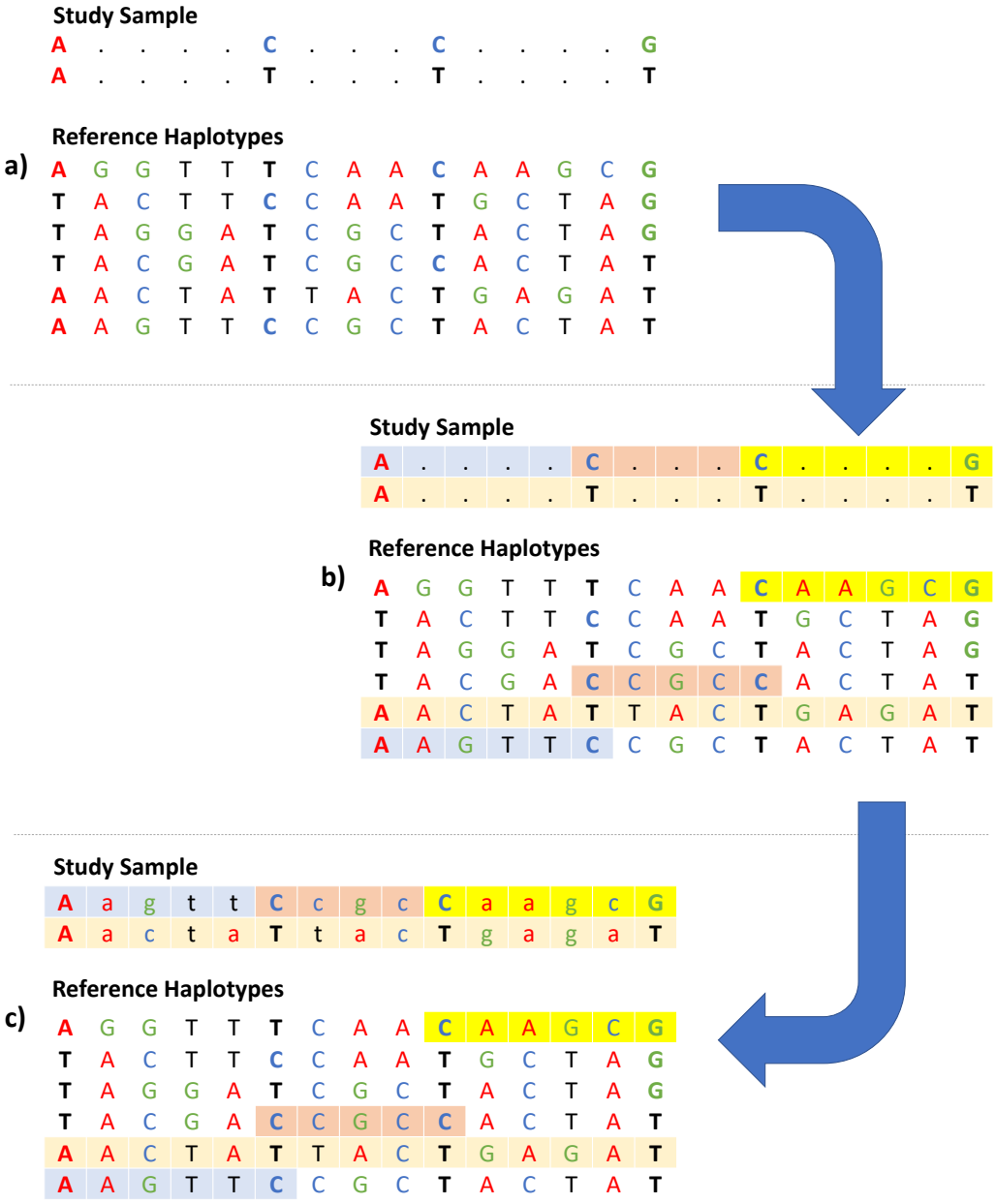
1. **Genotype imputation:**

   One of the important roles of phase information is its utilisation in estimating untyped SNPs (not determined experimentally) through learning from a reference panel of high-density haplotype data that contain the untyped SNPs (Das et al., 2018; Li et al., 2009; Marchini and Howie, 2010) as illustrated in the example shown in Figure 2.4. The majority of genotype imputation tools rely on phasing either during the imputation or "pre-phasing" of both the targeted data and the reference panel to reduce the computational cost dramatically (Browning et al., 2018; Das et al., 2016; Howie et al., 2012). The majority of imputation methods predict the alleles on each copy of a chromosome based on the similarity to the haplotypes within the reference panel as imputation is carried out for each copy separately. Therefore, phase information plays a major role with respect to this problem. In this context, haplotype information is obtained through computational phasing, and the accuracy of phasing influences the quality of the imputed genotypes (Loh et al., 2016b) used for several purposes such as the integration of several cohorts together when they have different SNPs, improvement of the resolution of published datasets, especially microarray-based genotypes, and fine-mapping of association findings (Das et al., 2018; Marchini and Howie, 2010). The rise of easily accessible genotype imputation as online services provided by both Michigan and Sanger imputation servers (Das et al., 2016; McCarthy et al., 2016) and the growth in terms of the number and size of datasets that need to be imputed has been a strong motivator to develop higher accuracy and scalability phasing tools. Around 50.7 million human genomes were prephased then imputed by Michigan imputation server in 2016-2020.

2. **Expression quantitative trait loci (eQTL) analysis:**

   Previous studies showed that the different allocation of the alleles on the chromosome copies can influence the regulation of some related genes in a different

**Figure 2.4: Example of genotype imputation.** A) An example of a study sample and reference haplotypes. The study sample contains alleles in four loci while the remaining alleles are missing. The reference haplotypes contain alleles for all loci including the ones in the study sample. Genotype imputation approaches rely on estimating the missing alleles in the study samples from the reference haplotypes. B) Genotype imputation tools identify shared regions between the study sample and reference haplotypes based on the alleles' similarity within the common loci. C) The missing alleles in the study sample are filled from the alleles in the shared region of the reference haplotypes (presented in lower case in the study sample). This Figure is inspired from Li, Yun, et al (Li et al., 2009).



manner (Tewhey et al., 2011; Ying et al., 2018). Figure 2.5 illustrates an example of the different impact of variation on gene expression. In this figure, there were three possible outcomes when having different allele combinations in two loci. The possible outcomes can vary more when considering additional SNPs. The impact of haplotype information on gene expression was investigated via expression quantitative trait loci (eQTL) analysis (Brown et al., 2017; Corradin et al., 2014; Garnier et al., 2013; Ying et al., 2018).

3. **Association with diseases:**

   It has been reported that including haplotype information in association analysis can potentially reveal genetics association that may otherwise be missed by keeping the focus only on genotype data (Browning, 2008; Kenny et al., 2009; Tregouet and Garelle, 2007; Trégouët et al., 2009; Wang et al., 2016), especially when targeting complex disorders (Liu et al., 2008) or when there is an association with rare variants (Howard et al., 2017). Genome-wide haplotype association studies have reported some haplotype associations with diseases such as Coronary artery disease (Trégouët et al., 2009), Major Depressive Disorder (MDD) (Howard et al., 2017), Acute Myeloid Leukemia (AML) (Lv et al., 2017), and Alzheimer's disease (AD) (Shang et al., 2015).

4. **Compound heterozygosity:**

   This term describes the case when two alternate alleles occur in a particular genomic region and each of these alleles is allocated on one homologous copy of a chromosome pair. Compound heterozygosity was reported for several disorders such as Severe Cardiac Conduction (Bezzina et al., 2003), Autism Spectrum Disorder (ASD) (Torres et al., 2018) and Skin Fragility-Woolly Hair Syndrome (Sprecher et al., 2004). Capturing compound heterozygosity requires knowledge of the location of alleles on the copies of each chromosome pair. This information is not available when considering the genotype of the SNPs or investigating the SNPs individually.

5. **Allele specific expression (ASE):**

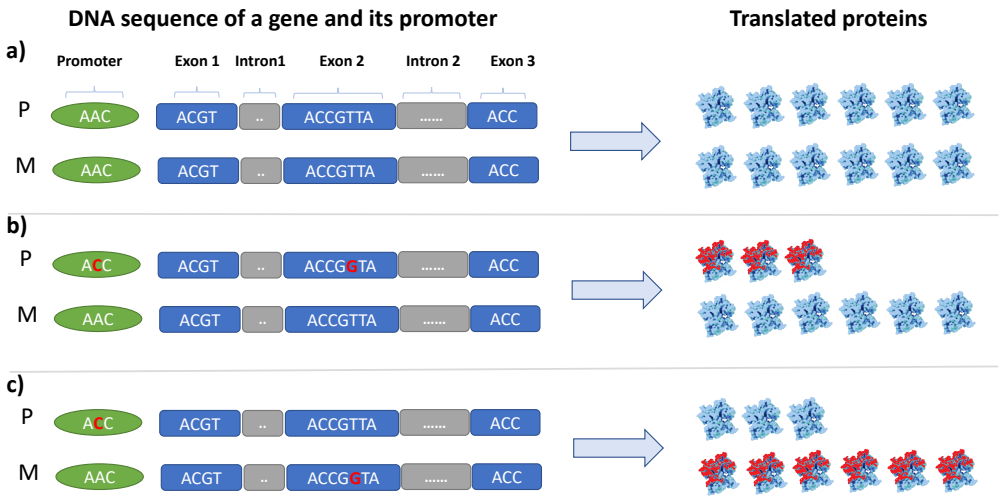
   It has been reported by various studies that there are some cases where one copy of a gene is expressed more than the other homologous copy (Pastinen, 2010; Tewhey et al., 2011). Cis-acting variations (allocated on the same copy with the affected

genes) within the regulatory elements of genes were reported to be associated with ASE along with other biological reasons (Buckland, 2004; Consortium et al., 2015; Li et al., 2006; Pastinen, 2010). Phase information can help to reveal alleles underlying ASE as it is important to consider the impact of alleles allocated on the same transcribed copy of a gene on the gene regulation.

6. **Fine mapping:**

   SNPs that are highly associated with diseases are not necessarily the real causal SNPs. They can be in high linkage disequilibrium with the real causal SNPs. Phase information is essential to link detected association with the real causal alleles when considering the extended haplotypes of the detected SNPs (Yang et al., 2008).

**Figure 2.5: The influence of allele location on gene functionality and regulation.** a) A case where there are no variations on both promoter and exons of a gene (assuming that both alleles do not impact the binding of RNA polymerase in the promoter or the structure of the translated protein). Both the regulation and the structure of the formed proteins are "as expected in a normal situation". b) and c): we assume that: 1- The allele "$C$" in the promoter region can affect the binding of RNA polymerase protein that initiates transcription process, therefore, having this allele will down-regulate the gene expression in the same homologous copy. 2- The allele in exon 2 affects the sequence of the gene, therefore, it affects the structure and functionality of the translated protein (affected protein is in red). b) Down-regulation of a nonfunctional gene in the paternal homologous copy, and normal regulation of the functional gene on the maternal copy when having the haplotype "$CG$" in the promoter and exon 2 in the paternal chromosome and the haplotype "$AT$" on the maternal one. c) Down-regulation of a functional gene in the paternal homologous copy, and normal regulation of the non-functional gene on the maternal copy when having the haplotype "$CT$" in the promoter and exon 2 in the paternal chromosome and the haplotype "$AG$" on the maternal one. This figure is inspired by (Tewhey et al., 2011).



### 2.3.4 Challenges when using haplotype information

Despite its significance, haplotype or phase information has not been used routinely in genetic studies, with genotype data the dominant information interrogated for genetics association (Buniello et al., 2018). The dearth of haplotype-based studies can be attributed to the limited availability of haplotype information and the increased complexity when dealing with the allocation of the alleles (See Figure 3.1 for the increased complexity issue). Below are the main challenges in more detail:

1. **Availability of haplotype data**:

   Most genome sequencing technologies determine the DNA sequence from a pool of paternal/maternal DNA fragments. The origin of these fragments is lost when the samples are mixed (Choi et al., 2018), therefore, it is hard to reassemble the sequenced reads into paternal and maternal copies for each chromosome. With these technical limitations, phase information is not readily available, but it can be estimated computationally via different approaches (will be explained later in this thesis, See section 2.4 Haplotype phasing). Using computationally estimated haplotypes increases the time and the efforts to conduct any study.

2. **Accuracy concerns**:

   Computational haplotype phasing is the main resource of data for haplotype-based studies. The accuracy of phased haplotypes is not maintained in long regions and the errors are not regularly distributed along the genome. It has been reported that it is almost impossible to phase a long region without an error (Browning and Browning, 2011; Marchini et al., 2006). In the experiments applied in this PhD, there was at least one error every 50 heterozygous SNPs within chromosomes 1, 6, and 17 (See Figure 3.4). There are many factors that influence the accuracy of phasing including the nature of the phased genomic regions and the characteristics of the used data. These issues are ignored with genotype-based analysis as genotype data are experimentally obtained and assumed to be very accurate (with respect to the errors that occurred during the genotyping procedure). Some concerns have been reported before with regard to using estimated haplotypes in genetic studies as real haplotypes (Curtis and Sham, 2006). The main concern is that estimated haplotypes with potential errors may lead to high type 1 errors

(False Positives). However, recent haplotype estimation or phasing tools have been reported to be very accurate, especially with the improved quality of genotype data such as increasing population size and density of SNPs (Browning and Browning, 2011; Loh et al., 2016a; O'Connell et al., 2016).

3. **Block determination**:

   The best investment of haplotypes is when considering them within blocks of multiple SNPs, not individual SNPs. Using haplotypes does not provide extra information when dealing with single SNPs compared to only considering the genotype of the SNP. Identifying the boundaries of haplotype blocks is not a straightforward procedure. Including extraneous loci or excluding important ones can affect the analysis substantially. There should also be a trade-off between the length of haplotype blocks and the accuracy of the phased blocks within the blocks as we demonstrate in chapter 3. Our experiments showed high error rate when using long blocks.

4. **Haplotype encoding**:

   The genotype of a SNP is encoded using alternate allele dosage, with 0, 1 and 2 are used to encode a SNP when it has 0, 1 and 2 alternate alleles respectively. Dosage encoding is simple, informative and constant, that is, the same encoding for a SNP across different datasets. It also can be considered as an additive (continues 0, 1 and 2) or categorical of three categories encoding which gives more flexibility when used with statistical models and tests (Shabalin, 2012). The flexibility of minor allele dosage encoding is not available with haplotypes within multiple SNPs. It is very hard to identify an additive impact of haplotypes within a block. The natural representation is the categorical encoding that cannot be constant as there is a high chance for a new haplotype to appear in different datasets (that can happen not only as a genuine novelty but also phasing error). This issue can limit the usability of pre-defined models such as regression or classification models trained on a specific dataset.

5. **Statistical concerns**:

   This issue arises when dealing with relatively small datasets (such as gene expression datasets). The increased number of possible haplotypes within a block has a negative impact on the statistical power of any model when investigating a few

samples. The genotype-based analysis does not suffer from this issue as there are only three possible values for any SNP as mentioned above.

## 2.4 Haplotype phasing

Haplotype phase can be determined via laboratory-based experimental methods or estimated from genotype data via computational and statistical methods termed haplotype phasing or estimation methods. Most haplotype phasing methods rely on extra information such as family or population genotypes to resolve an individual's genotypes into a pair of haplotypes. Population-based haplotype estimation is the most popular approach used for haplotype phasing as most available datasets are collected from populations of unrelated individuals.

Although the rapid development in genotype sequencing technologies that can be the main resource of haplotype information in the near future, the critical need for accurate computational haplotype phasing from genotype data can not be negated. Researchers want to interrogate large, existing genotype data gathered from thousands of individuals for several disorders. For example, UK Biobank [5] released recently genotype data for 500,000 individuals with more than 90 million SNPs, indels and large structural variants. European Genome-phenome Archive (EGA) [6] contains around 4,808 datasets for different populations and disorders. Computational haplotype phasing methods are needed to make use of existed datasets efficiently for different purposes such as disease heritability analysis, imputation of the untyped SNPs (Howie et al., 2012, 2009; Marchini and Howie, 2010), recombination analysis (Fearnhead and Donnelly, 2001; Stumpf and McVean, 2003) and population structure and history (Lawson et al., 2012).

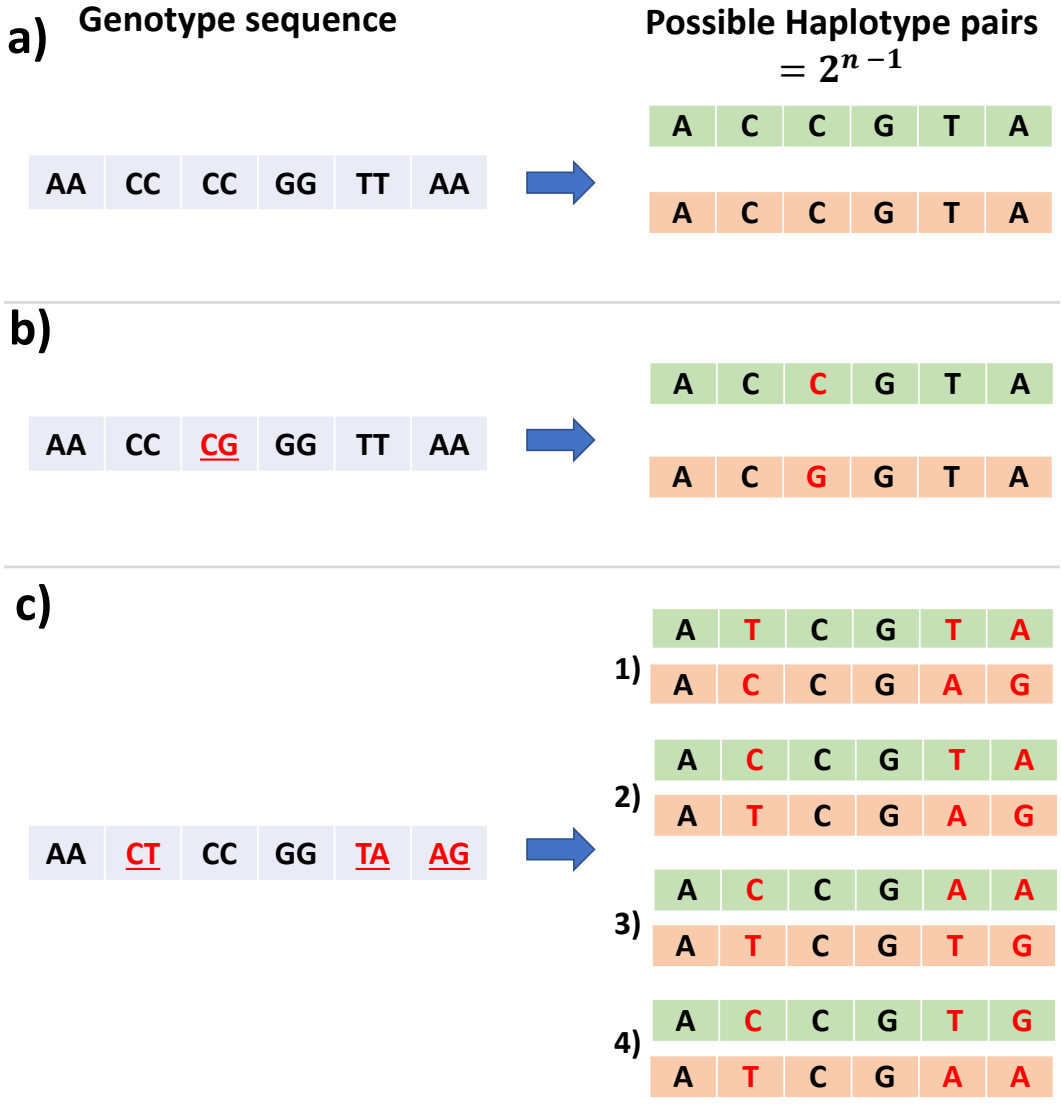
### 2.4.1 The problem of haplotype phasing

The most popular data representation of the DNA sequence is a genotypic sequence, where each SNP is represented as a combination of its alleles without considering their allocation on the paternal and maternal chromosome. The number of possible haplotype pairs of any genotype sequence increases exponentially with the number of heterozygous

[5] `http://www.ukbiobank.ac.uk/scientists-3/genetic-data/`
[6] `https://www.ebi.ac.uk/ega/datasets`

SNPs within the sequence. The simplest cases are when the heterozygous SNPs in the sequence are less than two. For such cases, the genotype sequence can be resolved by only one haplotype pair (See Figure 2.6 a and b). In general, there are $2^{n-1}$ potential haplotype pairs that can form a genotype sequence containing $n$ heterozygous SNPs. See Figure 2.6 (c) for an example of this general case.

**Figure 2.6: The possible haplotype pairs for a given genotype sequence.** **a)** The simplest case: there are two identical haplotypes for any genotype sequence of homozygous SNPs. **b)** The haplotypes of a genotype sequence containing one heterozygous SNP are identical at all loci with an exception to the locus of the heterozygous SNP where each copy has a different allele ("$C$" and "$G$"). **c)** A genotype sequence containing three heterozygous SNPs can be resolved by 4 possible haplotype pairs ($2^{n-1}$) formed by all possible combinations of the alleles within the heterozygous SNPs. Both a) and b) are unambiguous cases and can be resolved perfectly, however, the genotype sequence in c) can be estimated (potential errors) using extra information such as population or family genotype data.



### 2.4.2 First generation of the population-based haplotype phasing methods

Many haplotype phasing methods have been published in the last twenty years. The first generation of these methods fall into one of the following categories:

1. **Parsimony methods:**

   The first phasing approach was a parsimony approach introduced by Clark in 1991. Generally speaking, parsimony approaches aim to find the minimal haplotype set that can resolve a given genotype. The *Clark inference rule* that is the first parsimony method (Clark, 1990) consists of two stages:

   (a) Determine an initial haplotype set. The initial set can be pre-known haplotypes or resolved haplotypes of unambiguous genotype sequences (Contain fewer than two heterozygous SNPs).

   (b) Use the known haplotypes to resolve the genotype sequences in the data. A genotype sequence ($g$) can be resolved by a known haplotype sequence ($h_1$) and another haplotype sequence ($h_2$) that can form ($g$) with $h_1$. E.g. A given genotype sequence "*(CG)(CC)(TT)(AT)(AA)(GC)*" can be resolved by the known haplotype sequence "$GCTAAG$" and another complementary haplotype sequence "$CCTTAC$" (filling the alleles from the difference of genotype sequences alleles and the known haplotype alleles). The complementary haplotype sequence is added to the known haplotype set to be used in resolving other genotypes. The second stage is repeated until all genotypes are resolved or the known haplotypes cannot resolve any genotypes.

   Parsimony methods require an initial set of haplotypes to start and suffer from having many solutions or no solution at all (Salem et al., 2005). Clark suggested applying the method an enormous number of times by changing the order of the samples, then finding the solution that resolves most of the genotypes. Such approaches are suitable for a short region with no recombination events as it is very difficult to get the initial haplotype set to start the algorithm when trying to phase long regions. Examples of tools that used this approach are HAPINFERX (Clark, 1990) and HAPAR (Wang and Xu, 2003).

2. **Maximum likelihood methods:**

   These methods estimate the frequencies of the haplotypes that make the observed genotype data most likely. Most of them are based on Expectation-Maximization (EM) approach (Excoffier and Slatkin, 1995; Qin et al., 2002). With these methods, the haplotypes of the whole population are being estimated rather than estimating each individual separately. The accuracy of these methods increases with the population size. They are impractical for long regions as their space complexity increases exponentially with the heterozygous SNPs count (Browning and Browning, 2011). They also suffer from typical EM problems such as miscalling of low-frequency haplotypes and converging to non-global maxima (Salem et al., 2005). These approaches have been used in haplotype association studies (Tregouet and Garelle, 2007; Trégouët et al., 2009).

Both above approaches are not suitable for the current large datasets (the whole genome and thousands of individuals) as they are limited because of their space and time complexity, that increases exponentially to the heterozygous SNPs count.

### 2.4.3 Recent population-based haplotype phasing methods

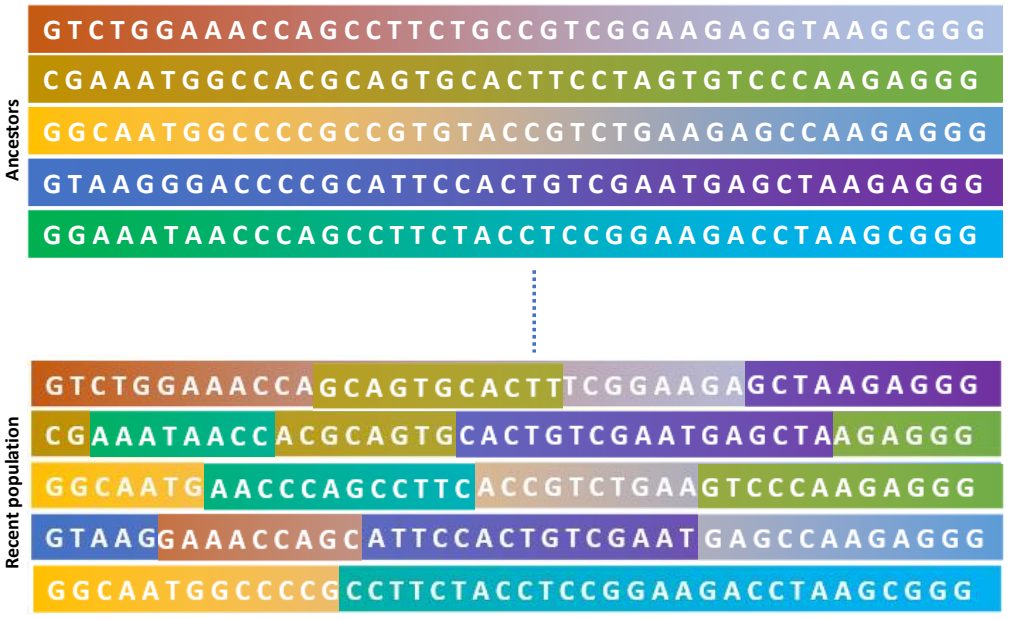
The majority of the current phasing methods use Hidden Markov Models (HMM) for haplotype phasing (Browning and Browning, 2007b; Howie et al., 2009; Li et al., 2010; Lin et al., 2004; Niu et al., 2002; Scheet and Stephens, 2006; Stephens et al., 2001). The main assumption of these methods is that the haplotypes of individuals are inherited from old generations yet they differ across time due to crossover, recombination and mutation events. The probability of a recombination event decreases in short regions that is why the haplotypes at a local scale with high linkage disequilibrium tend to be very similar among the individuals (Li and Stephens, 2003). The majority of recent haplotype phasing as well as genotype imputation methods are inspired by *Li and Stephens model* that was introduced in 2003 (Li and Stephens, 2003).

#### 2.4.3.1 Li and Stephens model

According to Li and Stephens model, there is a collection of a few different haplotypes within each contiguous genomic region due to the high LD. The complete haplotype

of an individual is formed from a combination of short haplotype stretches where each stretch is likely to be observed previously (Li and Stephens, 2003). Figure 2.7 illustrates the haplotype structure within a population according to the Li and Stephens model.

**Figure 2.7: The mosaic haplotypic structure of a population.** Each row represents the complete haplotype sequence of a random individual. The first half represents the ancestors of the individuals within the second half. The haplotype of the second individual in the bottom half of the population consists of five contiguous haplotype stretches identical to the stretches of individuals 2, 5, 2, 4, and 2 respectively within the same loci.




The main properties of Li and Stephen model are (Li and Stephens, 2003):

1. The haplotype of an individual is more likely to be similar to the most frequent haplotype within the population rather than the less frequent ones.

2. The chance to observe a unique haplotype within a region decreases with the number of unique haplotypes already observed within the region.

3. The chance to observe a unique haplotype within a region increases with the effective population size and the mutation rate within the region.

4. If the haplotype of an individual is not identical to a pre-existed one, it is very similar to it and not completely different.

5. The haplotype of an individual consists of different observed haplotype stretches with a length varying according to the recombination rate within the region. With

this assumption, the probability of observing a haplotype (can be computed using a hidden Markov model. See figure 2.8) is related to the probabilities of observing the consecutive haplotype stretches forming this particular haplotype.

The first two properties are similar to the assumptions of parsimony methods, as they also related unknown haplotype to pre-existed ones and try to minimise the number of unique haplotypes across a population.

#### 2.4.3.2 Examples of HMM-based haplotype phasing methods

Recent haplotype phasing methods share the same assumption as Li and Stephens model, however, they differ in the way they design their hidden Markov models such as creating the model's states from a single SNP or subset of consecutive SNPs, using random haplotypes templates or similar templates to the individual being phased. They also apply some optimisation to reduce the computation time such as investigating only the promising haplotypes and using an efficient data structure. These methods are accurate and suitable for large datasets (the length of the targeted region, and individuals count).

**PHASE (2000)** (Stephens et al., 2001) was considered the gold standard for phasing for a long time (Browning and Browning, 2011; Marchini et al., 2006). The time complexity of this algorithm increases exponentially with the heterozygous SNPs in the dataset which limits the practical use of this tool to small datasets. In the experiments applied in this PhD project, the execution time of PHASE when applied to a small dataset of 200 individuals and 140 SNPs exceeded ten days (the execution was terminated after this duration). It is reported that PHASE tool is only practical for up to 100 SNPs and a few hundreds of individuals (Browning and Browning, 2011). PHASE is not used anymore due to the large size of available datasets and the high accuracy of recent phasing tools. The first tool that was capable of phasing the SNPs at the genome-scale is **fastPHASE (2006)** (Scheet and Stephens, 2006). fastPHASE is the first modified version of the Li and Stephens Model (Das et al., 2018).

**IMPUTE (2009)** (Howie et al., 2009) and **MaCH (2010)** (Li et al., 2010) are both genotype imputation tools that do phasing implicitly. IMPUTE phases an individual's haplotype using a subset of the reference haplotypes chosen to be similar to the individual's haplotype, while MaCH uses a random subset of the random haplotypes. This

is how both tools improve the performance time compared to PHASE that uses all reference haplotypes.

**BEAGLE (2007)** (Browning and Browning, 2007b) uses localised haplotype cluster model similarly to fastPHASE but allows different genomic loci to have a different number of clusters (fastPHASE use fixed number of clusters). Among the previously mentioned tools, BEAGLE is the fastest and the most accurate for a population size $> 1000$ individuals. One of the factors that improved BEAGLE performance time is that BEAGLE does not consider all transitions between HMM states (at maximum two possible transitions for each state when considering bi-allelic SNPs, see Figure 2.8) as Li and Stephen model assumes as well all previous tools do (Browning and Browning, 2011). This constraint influences accuracy when BEAGLE is applied to small sample size dataset. The parsimonious model of BEAGLE requires to construct the model from all individuals' haplotypes rather than a subset of them as IMPUTE and MaCH does. MaCH is the slowest but the most accurate for small sample size (Browning and Browning, 2011).

**SHAPEIT2 (2012)** (Delaneau et al., 2012), **HAPI-UR (2012)** (Williams et al., 2012), **EAGLE2 (2016)** (Loh et al., 2016a), and **SHAPEIT3 (2016)** (O'Connell et al., 2016) are very popular and recent phasing tools. They are available on SANGER [7] and Michigan [8] imputation servers for pre-phasing. SHAPEIT3 is freely available for academic use only, and it was used to phase Haplotype Reference Consortium (HRC) dataset (McCarthy et al., 2016) that is used as a reference panel for genotype imputation in both mentioned servers. SHAPEIT3 is a recent version of SHAPEIT2 that is faster and more scalable than the old version but less accurate. EAGLE2 is a reference-based haplotype phasing tool, however, if there is no reference provided, it makes use of a long-range phasing tool EAGLE1 (Loh et al., 2016b) to create the initial set of haplotypes.

SHAPEIT2, SHAPEIT3, and HAPI-UR create HMM states for multiple SNPs instead of only one as other tools do. While the majority of the tools cluster similar haplotypes together to reduce the computational time, EAGLE2 considers all haplotypes as well as reduces the computation time by representing the haplotypes using an efficient data structure called *HapHedge* which is based on positional Burrows-Wheeler transform(PBWT) (Durbin, 2014).

[7] `https://www.sanger.ac.uk/science/tools/sanger-imputation-service`
[8] `https://imputationserver.sph.umich.edu/index.html`

SHAPEIT2, EAGLE2, BEAGLE are the most accurate tools among those mentioned tools according to the experiments applied in chapter 3. HAPI-UR is the fastest (but less accurate than the three mentioned tools), while BEAGLE is the slowest. SHAPEIT2 is not scalable for very large datasets $> 20,000$ individuals, while SHAPEIT3, EAGLE2, and HAPI-UR are. It is recommended to use SHAPEIT2 over SHAPEIT3 when the sample size is $< 20,000$ individuals (O'Connell et al., 2016). HMM models constructed for several approaches are illustrated in figure 2.8. More details about the development of haplotype phasing and more specifically, the tools mentioned above are available at following reviews (Browning, 2008; Browning and Browning, 2011; Klau and Marschall, 2017; Salem et al., 2005) and the supplementary materials of the study (Loh et al., 2016a).

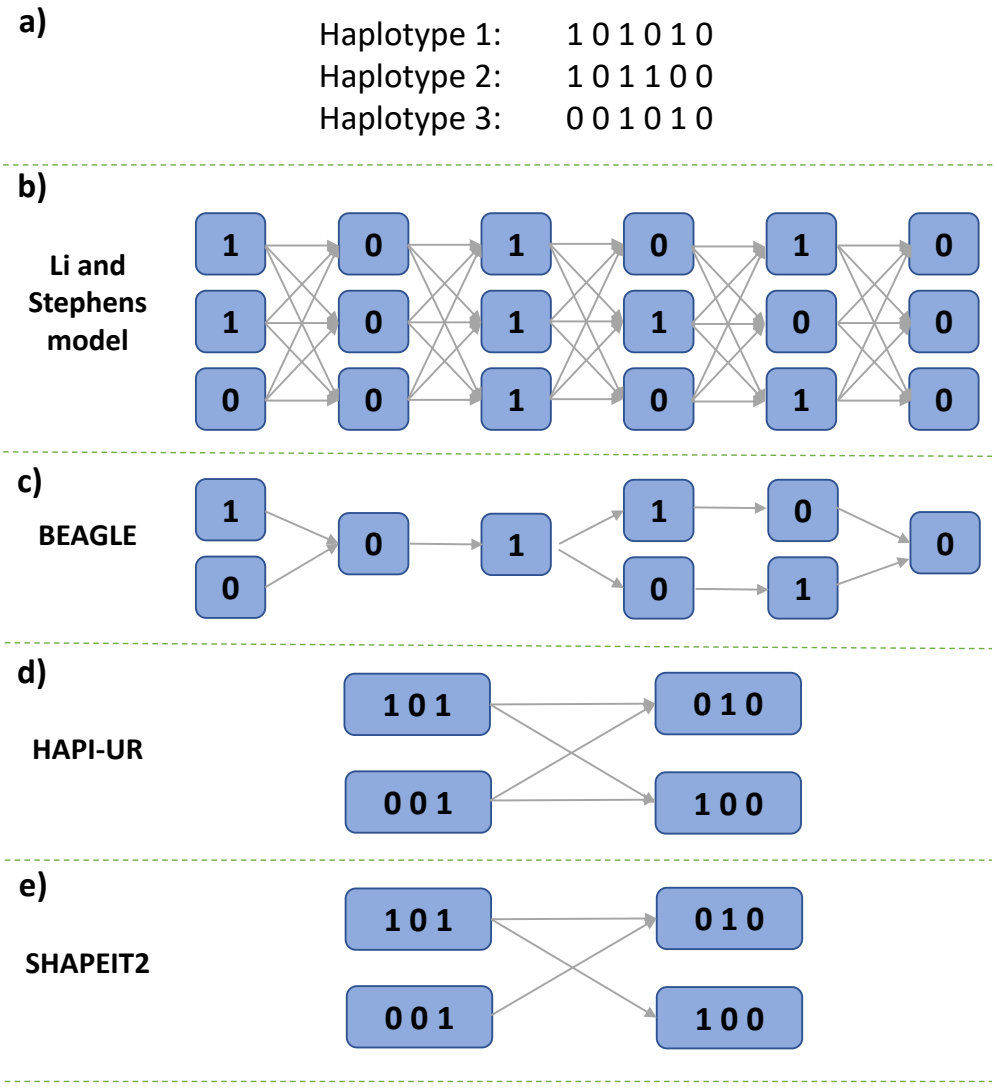
#### 2.4.3.3 Long range phasing

Long-range or identity by descent (IBD) phasing is a recent approach for haplotype phasing (Kong et al., 2008; Loh et al., 2016b). These approaches try to detect IBD regions that are haplotype segments shared between individuals and inherited from the same ancestor. Long-range phasing approaches exploit the detected IBD regions to accomplish the estimation using a combination of rule-based (similarly to parsimony methods) and HMM models. Such approaches can be used for related and unrelated individuals (Browning and Browning, 2011), however, they are more suitable for isolated populations with low immigration rates and hence less genetic heterogeneity. They also require to genotype a large proportion of the population to increase the chance of detecting IBD regions. A combined IBD and HMM approach used in EAGLE2 performed very well on different datasets (Choi et al., 2018; Loh et al., 2016a).

### 2.4.4 Family-based haplotype phasing

Family-based haplotype phasing is a specific case of related individuals phasing that is more similar to identity by descent (IBD) phasing. Here we explain resolving haplotypes for families (trios of a father, mother and child). When having genotype information of a family, haplotype phasing becomes a deterministic method. The principle of the family-based approach is Mendelian transmission law. A child inherits one allele from the father and another allele from the mother. Using this law, all child's heterozygous SNPs can be

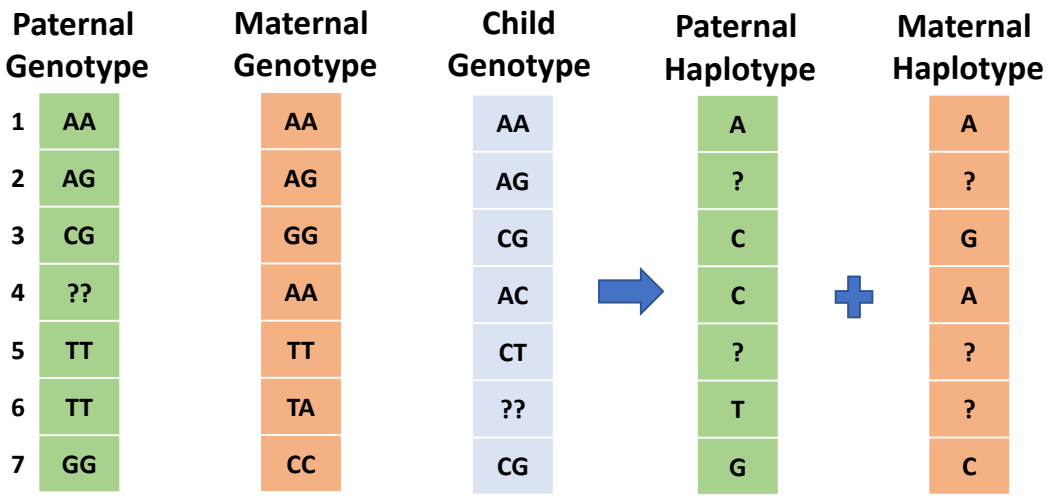
**Figure 2.8: An example of HMM models for the same reference haplotypes according to different phasing approaches.** a) Reference haplotypes used to create HMM. IMPUTE use subset of population haplotypes that are similar to the haplotypes of the individual being resolved. MaCH does the same but chooses a random subset. SHAPEIT2, SHAPEIT3, and HAPI-UR cluster similar haplotypes together (Loh et al., 2016a). PHASE, BEAGLE, and EAGLE2 use the haplotypes of all individuals. b) Li and Stephen model that is implemented by PHASE, fastPHASE, MaCH and IMPUTE (Browning, 2008; Williams et al., 2012). Each node represents a state (emitted allele or haplotype is written inside). Li and Stephen model allows the emission of all alleles at each state to account for mutations. c) BEAGLE's HMM built from an acyclic graph representation for haplotype cluster. This figure shows the optimised structure that leads to performance improvement compared to Li and Stephen model, however, there can be a situation where the model does not have haplotypes that can resolve a specific genotype sequence. d) and e) The states are created for haplotypes within a window of 3 SNPs.



resolved when one of the parents has a homozygous SNP in the same locus. See Figure 2.9 for a detailed example. A child's heterozygous SNPs cannot be resolved when both parents have heterozygous SNPs at the same locus. Using transmission law, most of the heterozygous SNPs can be resolved. For example, 80% of the child's heterozygous SNPs in 39 trios from the 1000 Genome Project could be resolved using this law in the experiments applied in chapter 3. The rest of the unresolved SNPs can be phased using other statistical approaches that exploit population data (explained above).

In general, the majority of genotype data is collected from unrelated individuals, therefore, this approach has limited practical usage. The most popular usage of family-based phasing method is to prepare gold standard haplotypes for evaluation purposes (Browning and Browning, 2011; Marchini et al., 2006). This usage is explained below in section 2.5.1 Haplotype data for evaluation.

**Figure 2.9: An example of using family information to determine the haplotype pair of a child.** The first three sequences from the left are the genotype sequences of the trio (father, mother and child). The paternal and maternal alleles at the first locus are determined simply as all SNPs are homozygous (most importantly the child). The child's heterozygous SNP at locus 2 cannot be resolved as both parents have heterozygous SNP at the same locus. The child's heterozygous SNP at the locus 7 can be resolved to "$G$" from the father and "$C$" from the mother as both parents have homozygous SNPs at the same locus. Similarly, the SNPs at the locus 3 and 4 can be resolved. "*??*" in the genotype sequence represent *missing genotype* which are unknown alleles due to genotyping errors. The inconsistent alleles of the trio's SNPs at the locus 5 represent a mendelian error.



### 2.4.5 Accuracy of haplotype phasing

Recent haplotype phasing tools have been reported to be very accurate (Loh et al., 2016a; O'Connell et al., 2016). However, these results were reported for recent datasets with large sample size and high SNP density. It is well-known that accuracy of phasing increases with the population size (Browning and Browning, 2011; Loh et al., 2016a; O'Connell et al., 2016) as almost all recent methods "learn" from other individuals' haplotypes and use them to resolve ambiguous haplotypes. The accuracy of phasing decreased 5 times when reducing the population size from 150,000 to 1,000 (Loh et al., 2016a; O'Connell et al., 2016).

Recent developments in this field target performance and scalability improvement (Loh et al., 2016a; O'Connell et al., 2016) which can be at the expense of accuracy (Browning and Browning, 2011). For example, SHAPEIT3 (O'Connell et al., 2016), a recent version of a well-known phasing tool (SHAPEIT2 (Delaneau et al., 2012)), is an example of this trade-off. It can handle large datasets that are impossible for SHAPEIT2 to process, yet it is less accurate and not recommended (by its authors) over SHAPEIT2 when dealing with small datasets (O'Connell et al., 2016). This direction of the development can be justified as there is also an indirect improvement of accuracy by including more individuals during phasing (more scalability). In other words, phasing a large dataset with the less accurate tool "may" lead to a similar accuracy to applying the more accurate tool on subsets of the data separately (if the accurate tool cannot phase the whole batch at once).

Even though the size and the resolution of genotype data are increasingly improving with the advent of new technologies in genome sequencing, most of the available datasets have been collected from small cohorts and low-density SNP markers (especially microarray data). For example, the mean sample size is 782, 691 and 2,463 samples of all datasets that contain more than 10 samples in European Genome-phenome Archive [9] (2,632 of 4,817 passed the filtration) and dbGaP data for general research use [10] array-based SNP genotypes (46 datasets) and NGS-SNP genotypes (17 datasets) respectively (Computed in July 2019). Unfortunately, available haplotype phasing tools do not phase such datasets with the same accuracy as when applied to high-quality datasets such as UK

[9] `https://www.ebi.ac.uk/ega/datasets`

[10] `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/collection.cgi`

Biobank data. There are few attempts to combine the outputs of several phasing approaches in a consensus estimator to improve the accuracy (Herzig et al., 2018; Williams et al., 2012). This consensus approach has been reported to be more accurate than one single approach, however, it is not well explored.

## 2.5 Evaluation of haplotype phasing

The need for accurate haplotype phasing for different purposes encouraged researchers to evaluate and compare the performance of haplotype phasing tools extensively (Browning and Browning, 2011; Choi et al., 2018; Herzig et al., 2018; Marchini et al., 2006; Miar et al., 2017; Salem et al., 2005). In this section, we explain the main issues related to haplotype phasing evaluation, that includes haplotype data used for evaluation, evaluation criteria and limitations of current evaluation studies.

### 2.5.1 Haplotype data for evaluation

One of the main obstacles of applying accurate evaluation of haplotype phasing that is observed in all reported evaluations is the limited availability of real haplotype data. As explained before in this thesis, real haplotype data is not readily available. This constraint made it hard to calculate accuracy metrics precisely when estimating haplotypes without having gold standard data for comparison. To overcome this issue, previous studies used the following datasets:

1. **Simulated data:**
   Haplotype data can be simulated using different tools such cosi2 (Shlyakhter et al., 2014), HapsiM (Montana, 2005), SimPed (Leal et al., 2005), and msprime(Kelleher et al., 2016). Simulated haplotypes can be randomly mated in pairs to form genotype data. The generated genotype data is phased then the estimated haplotypes are compared to the real ones used to form each genotype sequence. Such an approach was used in previous evaluations (Browning et al., 2018; Browning and Browning, 2007b; Li et al., 2010; Marchini et al., 2006; Stephens and Donnelly, 2003).

2. **Random mating of sex chromosome:**

   The popular form of genomic variation is genotype data representing a combination of paternal and maternal alleles. Males inherit one copy of chromosome X from their mother and copy of chromosome Y from their father. Therefore, those two chromosomes provide haplotype sequence as males have only one copy of those chromosomes. For evaluation purposes, the haplotypes of sex chromosomes can be randomly mated to form a genotype sequence then used in evaluation similar to the simulated dataset. This approach as used in previous studies (Scheet and Stephens, 2006).

3. **Family information:**

   In this scenario, a child's haplotypes are resolved using their parents' data and family-based phasing described in section 2.4.4 Family-based haplotype phasing. Parents genotypes are excluded while child data are combined with other unrelated individuals. The combined dataset is phased using population-based phasing tools and the estimated children haplotypes are compared to the real ones (resolved deterministically using parent's data). This partial evaluation indicates the quality of phasing for the whole dataset. This approach was used in the majority of the studies (Browning and Browning, 2011; Loh et al., 2016a; Marchini et al., 2006).

4. **Identity by descent:**

   IBD segments were used as accurate haplotypes for evaluation in a previous study (O'Connell et al., 2016).

### 2.5.2 Evaluation criteria

The evaluation of haplotype phasing concerns both accuracy and efficiency aspects of any tool. The efficiency of a tool can be measured via memory usage and performance time required to phase a dataset. There were several metrics used to evaluate the accuracy of phasing tools. We list below the most popular ones:

1. **Switch Error (SE):** This metric reflects phasing accuracy at the local scale and the errors are calculated according to neighbouring SNPs. It is considered the standard metric for haplotype phasing, and it is widely used since it was proposed in 2002 (Lin et al., 2002). Switch error is calculated at each individual scale as

follow:

$$SE = \frac{switches}{H - 1} \tag{2.4}$$

where *switches* is the number of the incorrectly phased heterozygous SNPs in comparison to their predecessor SNPs (SNPs in the previous genomic position), $H$ is individual's heterozygous SNPs count. An example of SE calculation is demonstrated in Figure 2.10. Switch accuracy can be calculated in the same way as:

$$SA = 1 - SE = \frac{H - 1 - switches}{H - 1} \tag{2.5}$$

2. **Incorrect haplotype percentage (IHP)**: The percentage of ambiguous individuals (their genotype sequence contains more than 1 heterozygous SNP) that are not perfectly phased (Clark, 1990; Stephens et al., 2001). It is an old metric that is only suitable for very short regions (a few SNPs). When the region is long enough, the incorrect haplotype percentage tends to 100% (Browning and Browning, 2011; Marchini et al., 2006).

3. **Incorrect genotype percentage (IGP)** The percentage of heterozygous genotypes (Missing SNPs can be included) that are incorrectly phased as a proportion of all heterozygous genotypes (Marchini et al., 2006). Contradictory to switch error, this metric reflects the quality of phasing at the whole sequence scale rather than locally. Most haplotype phasing applications are more concerned about phasing accuracy at a local scale which limits the usage of this metric (Browning and Browning, 2007b). The calculation of this metric requires to align each copy of the estimated haplotype pair to the most similar copy of the real ones. It is calculated as follows:

$$IGP = \frac{IH}{H} \tag{2.6}$$

Where $IH$ is the number of the incorrectly phased heterozygous SNPs and $H$ is Individual's heterozygous SNPs count. An example of IGP calculation is demonstrated in Figure 2.10.

4. **Missing error (ME):** Missing data or Missing SNPs are ambiguous SNPs that genotyping technologies couldn't determine their possible alleles, in other words, it is not known whether they are heterozygous or homozygous (minor or reference allele). Most studies apply some quality control procedures to eliminate such SNPs.

This metric is more related to genotype imputation evaluation, however, it can be used to evaluate phasing tools as most of phasing tools do genotype imputation for missing SNPs. The missing error is calculated as the percentage of missing SNPs that are incorrectly imputed as a proportion of all missing SNPs (Marchini et al., 2006). When measuring the missing error we only concern about the correctness of the imputation regardless whether the missing SNPs were correctly phased or not. The fact that missing SNPs are rare in most datasets, this metric has limited usage in this context. The equation of missing error calculation is:

$$ME = \frac{IM}{M} \tag{2.7}$$

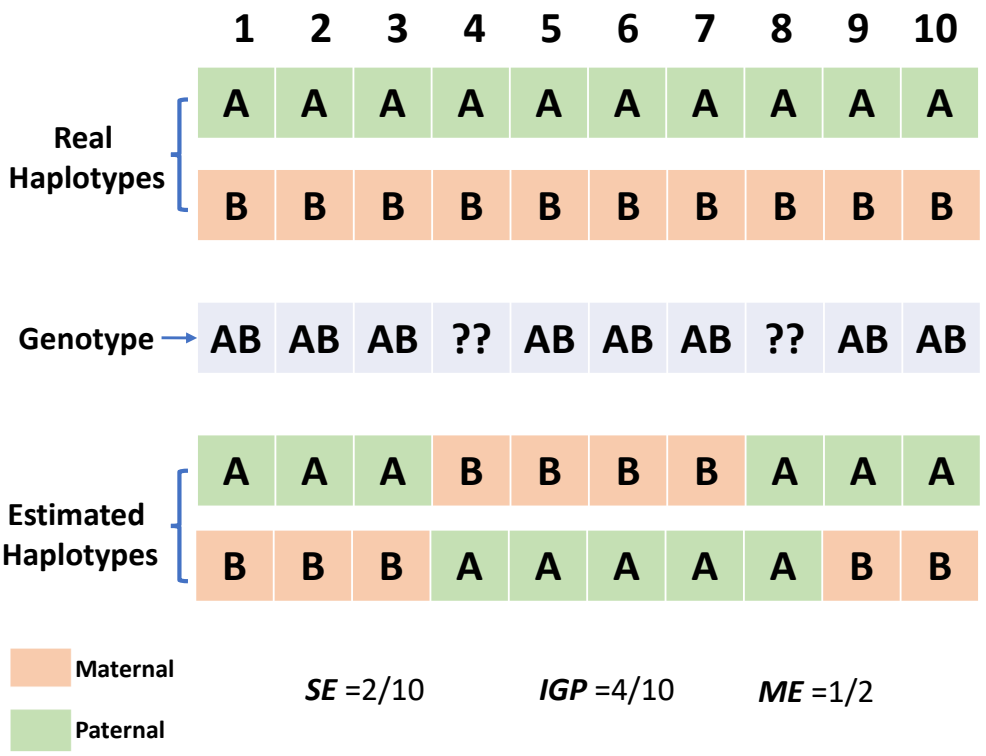
Where $IM$ is the number of the incorrectly imputed SNPs (heterozygous SNP imputed as homozygous and vice versa, homozygous SNPs imputed as a homozygous SNP but different alleles (minor allele instead of major allele and vice versa)) and $M$ is individual's missing SNPs. An example of ME calculation is demonstrated in Figure 2.10.

SE, IGP, and ME are calculated for each individual then summarised for the whole dataset. In the experiments applied in this PhD study, we used the mean of these metrics calculated for all individuals in the dataset. The equation of switch error calculation at population scale is mentioned in section 3.7.4 Evaluation criteria. There are other metrics used in this context such as comparing estimated haplotype frequencies to the population frequencies (Marchini et al., 2006), the percentage of correctly phased genes (Choi et al., 2018), and haplotype block length (Choi et al., 2018). Switch error is the dominant accuracy evaluation metric used in this field.

### 2.5.3 Factors impacting haplotype phasing

Despite the high accuracy reported for haplotype phasing tools, it is almost impossible to phase long regions without phasing errors (Browning and Browning, 2011; Marchini et al., 2006). The characteristics of both the dataset and the targeted genomic regions impact the accuracy of phasing. It has been reported that the accuracy of phasing increases with the population size, the density of SNPs, the quality of genotype data (less missing genotypes), and the relatedness of individuals (Browning and Browning, 2007a,

**Figure 2.10: Example of haplotype phasing evaluation metrics.** The sequences in this example are for 10 heterozygous SNPs as homozygous SNPs are not included in error calculation. There are 2 switches in the estimated haplotypes at loci 4 and 8 which led to 2/10 SE. When estimated haplotypes are aligned to the best matching real haplotype, there were 4 differences at loci 4, 5, 6, and 7. With respect to missing SNPs at loci 4 and 8, the latter is wrong as it is imputed as homozygous, therefore, ME is 1/2. As shown with the missing SNP at the $4^{th}$ locus, phasing error is not considered with ME calculation.




2011; Bukowicki et al., 2016; Marchini et al., 2006). Alleles frequency, recombination rate and individual ethnicity also impact on phasing.

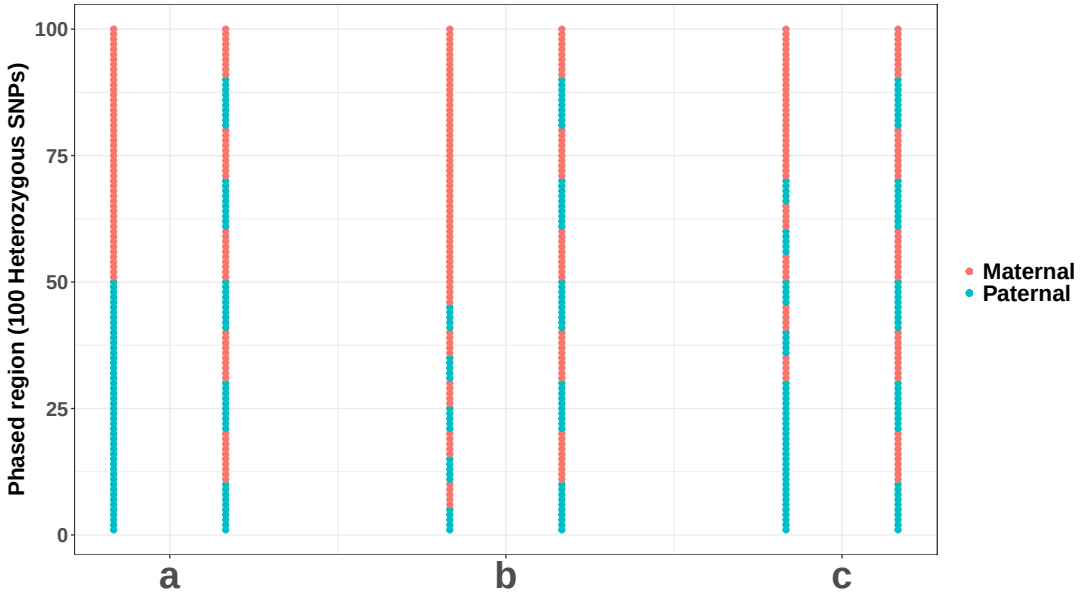
### 2.5.4 Limitations of phasing evaluation

Haplotype phasing was evaluated extensively on different datasets including large (Loh et al., 2016a; O'Connell et al., 2016) and medium to small (Browning and Browning, 2011; Marchini et al., 2006) human datasets and extended to cattle datasets (Miar et al., 2017). Phasing was assessed with respect to several influential factors such as population size (Browning and Browning, 2011; Loh et al., 2016a; O'Connell et al., 2016), minor allele frequency (Herzig et al., 2018), and missing SNPs rate (Herzig et al., 2018).

Unfortunately, all available evaluations do not take the application of phased haplotypes into account when assessing the accuracy. All accuracy evaluation metrics are summarised for each individual or the whole dataset. Such general perspective limits the ability to understand the consequences of phasing errors when conducting specific problems such as association analysis and genotype imputation; or dealing with regions that have special characteristics such as high/low recombination rate, LD or SNP density. It is not accurate to generalise the same error rate along the whole genome. For example, it is expected that the error rate increases in genomic hot recombination spots compared to regions with low recombination rate.

Even more, when two phased sequences have the same error rate (either switch error or incorrect genotype percentage), that does not mean they are phased similarly. Figure 2.11 illustrates three artificial ambiguous cases when switch error and genotype error have the same value for different estimations. The experiments applied in this PhD identified similar real cases. See 3.4.1 Different error locations obtained by different phasing tool and Figure 3.2. This ambiguity makes it hard to decide about the best metric reflecting the quality of the estimated haplotypes.

**Figure 2.11: Ambiguity of haplotype phasing evaluation metrics.** The figure represents two possible phases for the same region (100 heterozygous SNPs). The green SNPs are matching the paternal haplotype, while the red ones matching the maternal. a) different phases with different switch error (1 vs 5) but the same genotype error (50). b) Different phases with different genotype error (25 vs 50) but the same switch error (5). c) Different phases but the same switch and genotype error (5, 50)



It is very important to consider the context where phased haplotypes will be used and how they will be used to conduct an accurate assessment. For example, one of the most important applications of phased haplotypes is association analysis either with a disease (Howard et al., 2017; Lv et al., 2017; Shang et al., 2015) or gene expression (Brown et al., 2017; Garnier et al., 2013; Ying et al., 2018). In both cases, the genome is broken into short blocks or regions that are assessed statistically to reveal any significant associations. The typical evaluation metrics do not reflect the quality of phased haplotypes within these short blocks as the accuracy is not only influenced by the applied phasing tool but also the way the genome was partitioned. An extreme example to explain that is, there are definitely phasing errors when phasing applied to long regions. The incorrect haplotype percentage is 100% if the whole region is considered for association analysis without partitioning and regardless of which phasing tool is used. However, the incorrect haplotype percentage is 0% for the same phased region with the same errors if the whole region is partitioned into blocks of one SNP and that is also regardless which phasing tool is used. The latter part is conditioned by the assumption that there is no missing SNPs, or at least missing SNPs are correctly imputed. See 3.1 for another example.

Another issue to be considered is the impact of phasing on the downstream analysis. With respect to switch error, that is considered the standard metric for phasing evaluation and has been reported to be more informative than other metrics (Browning and Browning, 2007b), phasing tools were reported to be accurate by recent and old evaluations (the reported switch error in 2006, 2011, and 2016 is less than 5% for almost all tools). It is not hard to get high switch accuracy as haplotypes in short regions are very similar therefore, easy to be estimated correctly. What is really missing, is how the accuracy differences between different tools impact the downstream analysis. Is there a substantial gain when applying a phasing tool with 2% switch error compared to another one with 4% switch errors?. Furthermore, tools that have the same switch error (as a value) can have different errors when considering the location of the errors across the genome. The variation of the error locations can lead to different results when using phased haplotypes. To answer such questions, the evaluation should extend from a pure phasing evaluation to assess the impact of phasing errors on the accuracy of downstream analysis.

## 2.6 Haplotype block determination

Considering the haplotype at a SNP scale does not add extra information to genotype representation. The advantage of haplotype analysis is when considering the alleles in multiple loci. It has been reported that the human genome consists of consecutive blocks of different length, each of them containing a few different unique haplotypes (Gabriel et al., 2002; Zhang et al., 2002; Zhu et al., 2003). Several studies analysed these blocks for association analysis and fine mapping purposes (Zhang et al., 2002).

Several approaches were proposed to determine haplotype blocks. Below, We describe them in more detail.

1. **D-prime confidence interval**

   This approach is based on haplotype block definition proposed by Gabriel et al. (Gabriel et al., 2002). According to this study, a block can be determined from SNPs when there is a low chance of a recombination event to occur in the region containing the SNPs. The chance of having any recombination event within a region can be estimated from $D'$ between all SNP pairs within the region. Gabriel et al defined a haplotype block when the upper bound of 95% confidence interval is $> 0.98$ and the lower bound $> 0.7$. This approach is implemented via PLINK (Purcell et al., 2007) and haploview (Barrett et al., 2004) software with a possibility to reconfigure $D'$ parameters and the maximum distance between SNP pairs.

2. **Four gamete blocks**

   Another approach was proposed by Wang et al (Wang et al., 2002) to determine haplotype blocks when there is no recombination between any SNP pair. For each SNP pair $S_1$ and $S_2$ with alleles A, a and B, b respectively, observing the four gametes AB, Ab, aB, ab is an indication of having a recombination event between those two SNPs. Consecutive SNPs are grouped in one block if there is no recombination between them. This approach is implemented via haploview software (Barrett et al., 2004).

3. **Solid spine blocks**

   This approach is also based on $D'$ and implemented via haploview software (Barrett et al., 2004). A block can be determined for a group of consecutive SNPs as long

as the first and last SNP have high $D'$ (> pre-defined threshold) with each other and all SNPs in between. $D'$ between the inner SNPs (between the first and last SNP) is ignored and not considered in block determination (Barrett et al., 2004; Saad et al., 2018).

4. **Sliding window**

   Another popular approach to partition the genome into blocks is applying a sliding window. A sliding window scans the whole genome with/without overlaps and divide the genome into, in most cases, fixed-size blocks either based on SNP count (Trégouët et al., 2009) or recombination rates (Howard et al., 2017).

Haplotype block determination approaches were used extensively in association studies considering explicitly the phased haplotypes within the blocks (Howard et al., 2017; Lv et al., 2017; Shang et al., 2015; Wang et al., 2015; Wu et al., 2014) or the genotype of the SNPs within the block (Trégouët et al., 2009). Compared to the first three methods for block determination, the sliding window approach is less sensitive to the characteristics of the data or the region especially when the width of the window is determined by SNP count. However, it provides a more comprehensive scan as it can be applied with overlaps. It has been reported that four gamete blocks and D-prime confidence interval methods provide similar blocks compared to Solid spine blocks (Saad et al., 2018) that provides less but longer blocks.

An important factor to consider when choosing haplotype blocks with computationally phased haplotypes is the accuracy of haplotypes within the blocks. It has been reported previously that it is almost impossible to phase long regions without errors (Browning and Browning, 2011; Marchini et al., 2006), therefore, it is expected to have more errors in long blocks. The first three methods of haplotype block determination account for LD and recombination events when identifying the boundaries of the blocks. This leads to more accurate phased haplotypes within the blocks as phasing accuracy is correlated with these factors (Browning and Browning, 2007a). We report a comprehensive evaluation of phasing accuracy within the haplotype block determined by sliding window and confidence interval approaches in the third chapter of this thesis. More details about a comparison of the blocks produced by the first three approaches are described in this study (Saad et al., 2018).

## 2.7 Multiple test correction

A standard way to conduct association assessment such as SNP-gene expression, SNP-phenotype, or gene expression-phenotype involves two main steps:

1. Apply an adequate statistical test and calculate test statistics and p-value.
2. Compare the p-value with a significance threshold (typically $\alpha = 0.05$). Associations with a p-value less than $\alpha$ are considered significant, otherwise, they are not.

An adequate association assessment requires both, suitable statistical test, and suitable significance threshold. The choice of statistical test depends mainly on three factors. First, the purpose of the test, such as a comparison of groups or relationship assessment. Second, the type of the analysed data (categorical or quantitative). Third, the number of different groups in the data. Table 2.3 is a guideline for deciding about the adequate statistical test depends on the mentioned factors. When determining a fixed threshold $\alpha$, the chance for a few associations to have p-value $< \alpha$ purely by chance increases with the number of applied tests (McDonald, 2009). The determination of $\alpha = 0.05$ as a cutoff for significance determination means that the probability of incorrectly considering an association as significant is 5%. Converting this into numbers, there can be 5 incorrect findings when applying 100 tests, 50 when applying 1,000 tests and so on. These incorrect finding are termed *false positives* or *type 1 error*.

What makes this issue very important in the biological context is the substantial number of applied tests. In GWAS, all SNPs at a genome-wide scale are tested for association with a phenotype. Such analysis requires applying more than a million tests according to the SNP count in the used dataset. In eQTL analysis applied at the genome-scale, the possible test count is $n \times m$, where $n$ is the SNP count and $m$ is the gene count. To account for this issue, there have been several proposed strategies that adjust either the cutoff threshold or the p-value to reduce type 1 errors.

Multiple test correction is a general problem in statistics, especially when several statistical tests are applied simultaneously. In this section, the most popular approaches for multiple test correction will be explained in the context of association analysis. Regardless whether the association is between SNP-gene, SNP-phenotype or gene-phenotype.

1. **Controlling the familywise error rate: Bonferroni correction:**
   Controlling the familywise error rate means controlling the probability of having false positive or type 1 error, that is, controlling the value of significance cutoff $\alpha$. Bonferroni correction determines the significance threshold based on the number of applied tests. As $\alpha = 0.05$ is widely accepted in genetic problems, the corrected threshold according to Bonferroni correction is $\bar{\alpha} = \alpha/n$ where $n$ is the number of applied tests. For example, the new threshold for testing the association between 1,000 SNPs and a phenotype requires the application of 1,000 individual statistical tests, therefore, the corrected cutoff is $\bar{\alpha} = 0.05/1,000 = 0.00005$. Any association with a p-value $< 0.00005$ is considered significance with respect to Bonferroni correction.

2. **Controlling the false discovery rate: Benjamini-Hochberg procedure:**
   This approach aims at controlling the proportion of the false discoveries (incorrectly reported to be significant associations) of all discoverers (reported to be significant associations). The most popular procedure to control the false discovery rate is published by Benjamini and Hochberg in 1995 (Benjamini and Hochberg, 1995). According to Benjamini and Hochberg correction, significant discoveries are determined as follows (McDonald, 2009):

   (a) Determine the false discovery rate you accept for your tests. $\alpha = 0.05$ is widely accepted.

   (b) Calculate *p-value* for each test and sort them from smallest to largest.

   (c) Rank all tests based on their *p-value* from $i = 1$ *to* $m$, where $m$ is the number of tests. The smallest *p-value* has a rank $i = 1$.

   (d) Calculate Benjamini-Hochberg critical value for each $p\text{-}value_i$ as follow: $BH_i = (i/m) \times \alpha$. Where $i$ is the rank of the test according to its *p-value*, $m$ is the test count, $\alpha$ is the false discovery rate determined in (a).

   (e) Find the $p\text{-}value_i$ with the smallest $i$ that satisfies this condition: $p\text{-}value_i >= BH_i$.

   (f) All tests with a $p\text{-}value_i <$ the *p-value* determined in (d) are considered significant, that includes the one identified in (d).

3. **Permutation test**
   The principle behind the permutation correction is, discoveries that are reported

significant when a statistical test is applied to a dataset, should not maintain their significance when shuffling the samples within the dataset randomly (Camargo et al., 2008; Cheverud, 2001). The main steps to conduct this correction are:

(a) Calculate test statistic and p-value with respect to the original dataset.

(b) Shuffle the data randomly (shuffle the phenotype between samples, shuffle the gene expression values for eQTL analysis).

(c) Repeat the analysis in (a) in the shuffled/permuted dataset. Save the results.

(d) Repeat (b) and (c) large number of times ($n$). It has been recommended to repeat this analysis at least 1,000 times to estimate a 0.05 threshold of significance and 10,000 times to estimate a 0.01 threshold (Cheverud, 2001; Churchill and Doerge, 1994).

(e) Two types of comparisons can be carried out from the recorded results, either to estimate the threshold at each test (applied n times for n permutation) or the whole experiment scale (Churchill and Doerge, 1994). For example, the significance hypothesis is rejected for an association if its p-value with respect to the original dataset not in the 5% smallest p-values that are calculated for the random shuffles.

Both Bonferroni and Benjamini-Hochberg correction approaches assume the independency of the applied tests. If this assumption is violated, both approaches produce conservative thresholds (McDonald, 2009). Benjamini-Hochberg correction is preferred over Bonferroni when having a large number of tests, as the latter tends to give very conservative thresholds in this case (more false negatives) (Chen et al., 2017) (the threshold is divided by the number of tests). In general, Benjamini-Hochberg correction is less conservative than Bonferroni correction, and it is accused by obtaining more false positives (Camargo et al., 2008). In the biological context, Bonferroni correction is too conservative (Sul et al., 2015), as statistical tests are applied for a large number of SNPs or genes, and there is always a correlation between SNPs, especially in short regions. The permutation test is considered robust and more adequate for dependent tests (Cheverud, 2001; Sul et al., 2015) as it estimates the threshold directly from the data being analysed. However, it is computationally extensive due to the need to repeat the whole analysis numerous times (Sul et al., 2015).

Some standards thresholds are accepted to be used directly without applying multiple test correction. $5 \times 10^{-8}$ is a standard threshold in GWAS studies when applied to common SNPs (MAF$>$ 0.05) (Fadista et al., 2016). Other available corrections but "less common at least in biological context" are: Holm correction (Holm, 1979), Hochberg correction (Hochberg, 1988), and Hommel correction (Hommel, 1988).

**Table 2.3:** Guidelines for statistical test decision making.

| Purpose | Data | Test |
|---|---|---|
| Compare groups for significance difference | Categorical | Chi Square tests |
| Compare one group or a variable to a known value | Numerical | one sample t test |
| Compare two related groups or variables (normally distributed) | Numerical | paired sample t test |
| Compare two related groups or variables (not normally distributed) | Numerical | Wilcoxon signed rank test |
| Compare two unrelated groups or variables (normally distributed) | Numerical | independent samples t test |
| Compare two unrelated groups or variables (not normally distributed) | Numerical | Mann-Whitney test |
| Compare two or more groups (normally distributed) | Numerical | one-way ANOVA |
| Compare two or more groups (not normally distributed) | Numerical | Kruskal-Wallis |
| Identify significant relationships between variables | Ordinal or not normally distributed data | Spearman or Kendall's tau rank correlation |
| Identify significant relationships between variables | numerical and normally distributed data | Pearson correlation |
| Identify significant relationships between one dependent variable and two or more independent (predictor) variables | numerical data | Linear regression analysis |
| Identify significant relationships between one dependent variable and two or more independent (predictor) variables | numerical dependent variable and categorical independent variable | ANOVA |

## 2.8 Expression quantitative trait loci (eQTL)

The majority of genetic associations with a disease have been reported for variants allocated on non-coding genomic regions (Consortium et al., 2012b; Maurano et al., 2012; Nica and Dermitzakis, 2013; Schaub et al., 2012; Watanabe et al., 2017), which limit the biological interpretation of such associations. Therefore, these associations were either linked to other coding regions variants in high linkage disequilibrium with the detected associations (Spain and Barrett, 2015; Westra et al., 2018) or investigated to determine if they have an impact on the regulation of some associated genes especially when such variants are allocated on regulatory regions including promoters and enhancers (Nica and Dermitzakis, 2013; Zhang and Lupski, 2015). The impact of genomic variations on the level of gene expression was investigated through numerous studies called expression quantitative trait loci (eQTL) studies (Gilad et al., 2008; Mackay et al., 2009; Nica and Dermitzakis, 2013). It has been reported that more than 50% of SNPs revealed by GWAS studies in the brain are associated with gene expression levels (Gådin et al., 2019; Jaffe et al., 2018).

eQTL analysis can be applied locally, that is, pairs are formed from SNPs and genes that are close to each other physically (genetic distance) or globally through a comprehensive scan of all possible SNP-gene pairs at genome-wide scale (Tian et al., 2014). With respect to these two applications, significant SNP-gene association can be classified into *cis* association when the SNP and gene are close to each other, or *trans* when the distance between the SNP and the associated gene is relatively large. The majority of previous studies used a distance of 1Mb up/downstream of the associated gene as a cutoff for the distinction between cis and trans associations (Nica and Dermitzakis, 2013; Tian et al., 2014). Other cutoffs used by other studies are 100 kb (Pickrell et al., 2010), 200 kb (Montgomery et al., 2010) or 2.5 Mb (Kreimer and Pe'er, 2013).

### 2.8.1 eQTL analysis methods

The popular approach to conduct eQTL analysis is to apply pairwise correlation or regression analysis across SNP-gene pairs (Tian et al., 2014). Different statistical and correlation assessments can be applied in this context, such as Pearson correlation, Spearman rank correlation, linear regression or analysis of variance (ANOVA) (Shabalin,

2012), then a p-value is reported for each pair of SNP-gene. Significant associations are identified after the application of multiple test correction and the determination of significance threshold. The choice of the statistical test used in eQTL analysis is dependent on the assumption whether the SNP is a nominal variable or continues variable with respect to the minor allele dosage. When a SNP assumed to have an additive impact on the gene expression, a simple linear regression model can represent this linear relation as illustrated in equation 2.8. With a simple linear regression model, a model representing the relationship between two continues variables, the common test statistics: $t$, $F$, R-squared $R^2$ and likelihood-ratio $LR$ are equivalent and can be calculated based on the sample correlation coefficient or sample Pearson's correlation coefficient $r$ (Shabalin, 2012).

Consider a population of $n$ individuals with known genotypes and gene expression, the association between any SNP-gene pair can be assessed using one of the following statistical methods:

#### 2.8.1.1 Regression analysis

The level of any gene expression can be expressed as a function of the minor allele dosage of a SNP as illustrate in equation 2.8

$$g = \alpha + \beta s + \epsilon \tag{2.8}$$

Where $s$ is the dosage of the minor allele within the SNP , $s \in (0, 1, 2)$. Based on this model, $F$ statistics and p-value of significance can be calculated by carrying out the following steps:

1. Calculate **Corrected Sum of Squares for Model**

$$SSM = \sum_{i=1}^{n} (\hat{g}_i - \bar{g})^2 \tag{2.9}$$

   where $\hat{g}$ is the regressed gene expression and $\bar{g}$ is the mean of the gene expression.

2. Calculate **Sum of Squares for Error**

$$SSE = \sum_{i=1}^{n} (g_i - \hat{g}_i)^2 \tag{2.10}$$

where $g$ is the actual value of the gene expression.

3. Calculate **Corrected Sum of Squares Total**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.11}$$

4. Calculate **Mean of Squares for Model**

$$MSM = SSM/DFM \tag{2.12}$$

where $DFM$ is the corrected degrees of freedom for model, that can be calculated as $DFM = p - 1$ where $p$ is the number of predictors or coefficients in the model. In simple linear regression, the case when assessing an association between single SNP and a gene expression level, $DFM = 1$.

5. Calculate **Mean of Squares for Error**

$$MSE = SSE/DFE \tag{2.13}$$

where $DFE$ is the corrected degrees of freedom for model, that can be calculated as $DFE = n - p$.

6. Calculate **F-test**

$$F = \frac{MSM}{MSE} = (SSM/DFM)/(SSE/DFE) \tag{2.14}$$

7. Calculate **p-value** based on F-test

$$p\text{-}value = 2 \times pt(-abs(F), DFM, DFE) \tag{2.15}$$

#### 2.8.1.2 Correlation assessment

Pearson correlation is used to assess the association between two continues variables, then t-test and p-value is calculated from the correlation value $r$.

$$r_{(g,s)} = cor(g, s) = \frac{\sum (s_i - \bar{s})(g_i - \bar{g})}{\sqrt{\sum (s_i - \bar{s})^2 \sum (g_i - \bar{g})^2}} \tag{2.16}$$

Where $g_i$, $s_i$ are gene expression and SNP value (0, 1 or 2) for an individual $i$, respectively. $\bar{g}$ and $\bar{s}$ are the mean of the gene expression and the SNP, respectively. t-test can be calculated from the correlation as follow:

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \tag{2.17}$$

Where $n$ equals to the sample count or the individual count in the dataset. p-value is calculated through t-test distribution function with $n-2$ degrees of freedom, where $n$ in individual count in the dataset.

$$p\text{-}value = 2 \times pt(-abs(t\text{-}statistic), df = n-2) \tag{2.18}$$

As mentioned before, all the following statistical tests can be calculated from the correlation coefficient $r$:

$$F = t^2 = (n-2)\frac{r^2}{1-r^2} \tag{2.19}$$

$$R^2 = r^2 \tag{2.20}$$

$$LR = -n\log(1-r^2) \tag{2.21}$$

#### 2.8.1.3 Analysis of variance (ANOVA)

Analysis of variance (ANOVA) model can be used to represent the relation between gene expression and a SNP when a SNP is considered to have a categorical value. ANOVA model can be represented as follow:

$$g = \alpha + \beta_1 s_1 + \beta_2 s_2 + \epsilon \tag{2.22}$$

Where $s_1 = I(s = 1)$ and $s_2 = I(s = 2)$. Based on this model, $F$ statistics can be calculated from $R^2$ as follow (Shabalin, 2012):

$$F = \frac{(n-k-1)R^2}{k(1-R^2)} \tag{2.23}$$

Where $k = 2$ for having two predictors: $s_1$ and $s2$, and $R^2 = r^2_{(g,s_1)} + r^2_{(g,s_2)}$ as defined in 2.16. p-value can be calculated from F-test.

#### 2.8.1.4 Lasso-based models

In a typical eQTL analysis, associations are investigated for a large number of SNP-gene pairs and for small number of individuals (relatively). Features selection and spare learning models when considering the high dimensional nature of both gene expression ($k =\sim 20,000$ genes) and SNPs ($j > 200,000$ SNPs) where both dimensions are grater than individual count ($n < 1,000$ usually) and the assumption that there are few associations between SNPs and gene expressions (Chen et al., 2012; Tian et al., 2014). The relation between gene expression is SNPs can be expressed via the equation:

$$G = SB + E \tag{2.24}$$

where G is a matrix of gene expression levels for $k$ genes of $n$ individuals as illustrated below:

$$G = \begin{bmatrix} g_1^1 & \dots & g_1^k \\ \vdots & \ddots & \vdots \\ g_n^1 & \dots & g_n^k \end{bmatrix} \tag{2.25}$$

S is a matrix of $j$ SNPs of $n$ individuals as illustrated below:

$$S = \begin{bmatrix} s_1^1 & \dots & s_1^j \\ \vdots & \ddots & \vdots \\ s_n^1 & \dots & s_n^j \end{bmatrix} \tag{2.26}$$

A standard multi-task Lasso model (Least absolute shrinkage and selection operator) (proposed in 2.28) aims to find the optimal association coefficients matrix $B$ (see 2.27) by minimising square loss function and a regularization term while it accounts in the same time for correlation between the expression of different genes (Tian et al., 2014).

$$B = \begin{bmatrix} \beta_1^1 & \dots & \beta_1^k \\ \vdots & \ddots & \vdots \\ \beta_j^1 & \dots & \beta_j^k \end{bmatrix} \tag{2.27}$$

$$min_B \sum_{k=1}^{K} \parallel g^k - S\beta^k \parallel_2^2 + \lambda \sum_{j=1}^{J} \delta_j \parallel \beta_j \parallel_2 \tag{2.28}$$

Where $g^k$ is the gene expression of the gene ($k$), $\beta^k$ represents the association coefficients for $J$ SNPs with the expression levels of the gene $k$. In a minimised $B$, there are a small number of non zero coefficients that represent the significant eQTL associations.

### 2.8.2 Haplotype-based eQTL analysis

Similarly to other genetic problems, the use of haplotype information in eQTL analysis was also limited to very few studies (Brown et al., 2017; Corradin et al., 2014; Garnier et al., 2013; Ying et al., 2018). Reported results demonstrate the efficacy of including alleles allocation into account to reveal significant associations between genetic variations and gene expression that cannot be captured by ignoring phase information. For example, four SNPs within two enhancers have been reported to have combined haplotypic impact on gene expression regulation while separately they do not have this impact (Corradin et al., 2014). In another study, 2,650 RNA probes out of 19,805 (13.4%) have been reported to be associated with haplotypes allocated within 200 kb from the start and end of each prob. The p-value of these associations where $< 10^{-4}$ the p-value of any single SNP within these regions (Garnier et al., 2013).

Phase information used within these studies is either provided from the source (GTEx and 1000 genome project) as mentioned in both studies (Brown et al., 2017; Ying et al., 2018) or obtained computationally using Expectation-Maximisation (EM) inference (Garnier et al., 2013). EM phasing was applicable in this study as it was applied to relatively short regions 200 kb surrounding each RNA prob. For longer regions, utilising this approach can be problematic due to its high space complexity. This constraint can be eliminated by using current haplotype phasing approaches mentioned in the background as they are not only more accurate but also faster and more scalable than EM algorithm.

Previous haplotype-based eQTL studies have suggested interesting solutions for haplotype determination approach which is one of the main challenges when conducting haplotype-based analysis (as described in the background at the section: 2.3.4 Challenges when using haplotype information). Some studies limited the determination of haplotypes to regulatory elements such as enhancers and promoters (Corradin et al., 2014; Ying et al., 2018). Such a focus is motivated by the fact that many of eQTL findings were reported for SNPs allocated on the regulatory regions (Maurano et al., 2012;

Nica and Dermitzakis, 2013). In the study (Ying et al., 2018), haplotypes were formed from SNPs allocated on one regulatory region through an iterative process. Haplotypes are saved from one SNP, then another SNP is added and the haplotypes are extended to two SNPs and saved, and so on. The haplotypes keep extending until the number of haplotypes exceeds a predefined threshold (10,000 in this study). When the threshold is reached the scanned window is determined from the first SNP to the last added SNP. The same iterative procedure is applied starting from the SNP in the SNP in the centre of the previous scanned window. Haplotype determination is done for each gene in 1mbp up/downstream of the gene. Haplotypes are filtered based on their frequency ($< 0.05$ are eliminated) then assessed using linear regression. While the study (Corradin et al., 2014) also focuses on regulatory elements (mainly enhancers) the same as the previous study (Ying et al., 2018), it considers haplotype within multiple variants allocated on different enhancers that are targeting the same genes.

Another study used haplotype information to encode two SNPs to account for the cases where the minor allele is allocated on one copy of the paternal/maternal haplotypes or both, regardless of which SNPs(Brown et al., 2017). For example, the encoding of a pair of SNPs will be 0 if there is no alternate allele within the pair of SNPs. 1 if there is at least one alternate allele on either the paternal or maternal haplotypes but not on both. 2 when there is at least one alternate allele on each of the paternal and maternal haplotypes. SNP pairs were investigated within 1mbp surrounding the target gene (Brown et al., 2017). Considering all possible pairs within 1mbp regions is not only computationally expensive but also introduce very conservative significance cutoffs especially after multiple test correction. Therefore, the authors limited the SNP pairs to the ones that pass 0.4 significance threshold after Bonferroni correction and when the each SNP of the pair is not highly correlated with the pair's encoding mentioned above ($r < 0.8$ the threshold used in their study).

Another haplotype-based eQTL analysis was applied to all possible haplotype combination of 1 to 4 "special" SNPs termed haplotype tagging SNPs (htSNPs) (Garnier et al., 2013). Haplotype tagging SNPs are determined as the most informative SNPs within a sliding window of ten consecutive SNPs when they represent $> 95\%$ of the possible haplotypes within this window. The fact that this study focused on associations with RNA probes, it was possible to investigate all possible combinations of 1 to 4 htSNPs (maximum 187 htSNPs as reported in this study). This investigation can be problematic

if eQTL analysis was conducting for each gene when considering all SNPs within 1 or 0.5 mbp window surrounding the transcription start site as the majority of eQTL studies do.

These studies lack the investigation of the potential impact of phasing accuracy on the applied eQTL analysis. It is well-known that phasing accuracy is maintained within short regions(Browning and Browning, 2011; Marchini et al., 2006). We demonstrate in the third chapter of this thesis (3.4.3.1 An evaluation of haplotypes obtained by a sliding window) and Figure 3.10 how the switch error increases within long regions. Such issue can significantly impact the findings of these studies, with the least impact is on the study (Ying et al., 2018) as haplotypes are considered only within regulatory regions that are relatively short (the average size is 1.25 kb as reported in the supplementary materials of the study).

Another missing information in this context is how the results will vary when considering the combined genotype of the multiple SNPs instead of their haplotypes. The evaluation and comparison reported were only limited to the single SNP results studies (Brown et al., 2017; Corradin et al., 2014; Garnier et al., 2013). It is very important to evaluate against the combined genotypes of multiple SNPs to understand and determine whether the difference in the reported results is only related to the consideration of multiple SNPs together (instead of separately) or also to the consideration of the allele allocation in the analysis.

The use of haplotype information in genetic association studies such as genome-wide haplotype association studies (GWHAS) (Howard et al., 2017; Lv et al., 2017; Tregouet and Garelle, 2007; Trégouët et al., 2009) or eQTL has been demonstrated to be a complementary analysis to single SNP-based analysis (Brown et al., 2017; Browning and Browning, 2008; Lambert et al., 2013). Such analysis was limited previously for several potential reasons that are explained in the background of this thesis. However, there is a need to explore and use such approaches to investigate the genetic association with diseases, especially with the diversity and the improved quality of genetic data as well as the improved accuracy of haplotype phasing tools that can provide accurate computationally phased haplotypes.

# Chapter 3

# Exploring effective approaches for haplotype block phasing

*This chapter contains the investigations performed to respond to the first research question:*

> *What are the limitations and the expected phasing error within the haplotype blocks determined either via sliding window or based on Linkage Disequilibrium(LD) when phased by computational haplotype estimation tools?*

*And partially to the second one:*

> *How can we maximise phasing accuracy by aggregating multiple phased haplotypes into a consensus estimator applied to datasets with different characteristics*

This chapter appears in a manuscript published in BMC Bioinformatics journal [1].

[1] Al Bkhetan, Ziad, et al. "Exploring effective approaches for haplotype block phasing." BMC Bioinformatics 20.1 (2019): 540.

## 3.1 Motivation

We have shown in the background that genetic studies concerning haplotype information investigate multiple SNPs together (in pairs or grouped in blocks) as considering haplotype information at individual SNP does not add extra information to the genotypic representation of the SNP. In addition, we explained the limitations of available evaluations of phasing that ignore the downstream application of phasing as well as the way phased haplotypes are used. Therefore, it is essential to understand and quantify the error rate within haplotype blocks that are the most popular way to consider haplotype information in genetic problems (especially linkage analysis). The accuracy of phasing within blocks is influenced by both the accuracy of haplotype phasing tools as well as the partitioning procedure used to divide individuals' genomes.

Furthermore, recent phasing tools have been reported to be accurate, however, there is no consensus agreement on a specific tool to outperform others all the time. While the switch errors reported for different tools applied to the same dataset are similar, it is not known whether the tools fail in the same locations or not. The investigation of error similarity or differences across different tools helps to make the best use of them.

In this chapter, we explore the errors obtained by different phasing tools and we conduct a preliminary analysis to assess accuracy improvement obtained by aggregating the outputs of phasing tools into a consensus haplotype estimator. Observations noted for the error patterns, as well as tools' instability, inspired the further evaluations and explorations of the efficacy of the consensus approach in the fourth chapter. In addition, the comparison of haplotype block determination methods is effectively employed in the development of haplotype-based eQTL method introduced in the fifth chapter. While we explore in this chapter the accuracy of phasing in the context of association analysis, there is no direct assessment of the impact of phasing on downstream analyses. Therefore, we complete this investigation by evaluating the impact of phasing errors on two different major applications of phased haplotypes, genotype imputation in chapter 4 and eQTL analysis in chapter 5. We believe that findings and observations obtained in this chapter can be applied to other haplotype-based genetic investigations such as GWAS.

Al Bkhetan *et al. BMC Bioinformatics* (2019) 20:540
https://doi.org/10.1186/s12859-019-3095-8

BMC Bioinformatics

RESEARCH ARTICLE Open Access

# Exploring effective approaches for haplotype block phasing

Ziad Al Bkhetan[1], Justin Zobel[1], Adam Kowalczyk[1,2,3,4], Karin Verspoor[1*†] and Benjamin Goudey[4,5†]

## Abstract

**Background:** Knowledge of phase, the specific allele sequence on each copy of homologous chromosomes, is increasingly recognized as critical for detecting certain classes of disease-associated mutations. One approach for detecting such mutations is through phased haplotype association analysis. While the accuracy of methods for phasing genotype data has been widely explored, there has been little attention given to phasing accuracy at haplotype block scale. Understanding the combined impact of the accuracy of phasing tool and the method used to determine haplotype blocks on the error rate within the determined blocks is essential to conduct accurate haplotype analyses.

**Results:** We present a systematic study exploring the relationship between seven widely used phasing methods and two common methods for determining haplotype blocks. The evaluation focuses on the number of haplotype blocks that are incorrectly phased. Insights from these results are used to develop a haplotype estimator based on a consensus of three tools. The consensus estimator achieved the most accurate phasing in all applied tests. Individually, EAGLE2, BEAGLE and SHAPEIT2 alternate in being the best performing tool in different scenarios. Determining haplotype blocks based on linkage disequilibrium leads to more correctly phased blocks compared to a sliding window approach. We find that there is little difference between phasing sections of a genome (e.g. a gene) compared to phasing entire chromosomes. Finally, we show that the location of phasing error varies when the tools are applied to the same data several times, a finding that could be important for downstream analyses.

**Conclusions:** The choice of phasing and block determination algorithms and their interaction impacts the accuracy of phased haplotype blocks. This work provides guidance and evidence for the different design choices needed for analyses using haplotype blocks. The study highlights a number of issues that may have limited the replicability of previous haplotype analysis.

**Keywords:** Haplotype estimation, Phasing, Haplotype blocks, Haplotype analysis

## Background

Most genetic studies focus on analyzing genotypes to detect significant genetic associations with diseases [1]. However, it has long been recognized that some disease-associated haplotypes, the specific allele sequence on each copy of homologous chromosomes, may be undetectable with a focus on genotype alone [2, 3]. For example, the different allocation of specific alleles on each copy of a chromosome pair (which is ignored by genotype analysis) can impact the gene expression of an associated gene in a different manner [2]. Phasing, also known as haplotype estimation, reconstructs the haplotype sequences from genotype data and has been essential for understanding sequence-specific variation such as allele-specific expression [4, 5], methylation effects [6], and compound heterozygosity [2, 7]. Moreover, numerous studies have shown that haplotype-based association analysis can identify variants that would be missed by a standard single nucleotide polymorphism (SNP)-based analysis [8–10].

Despite its promise, phased haplotype association analysis is not commonly applied in genome-wide association studies, likely due to the increased complexity of haplotype analysis. As shown in Fig. 1a, the analysis requires

*Correspondence: Karin.Verspoor@unimelb.edu.au
†Karin Verspoor and Benjamin Goudey contributed equally to this work.
[1]School of Computing & Information Systems, University of Melbourne, 3010 Parkville, Australia
Full list of author information is available at the end of the article

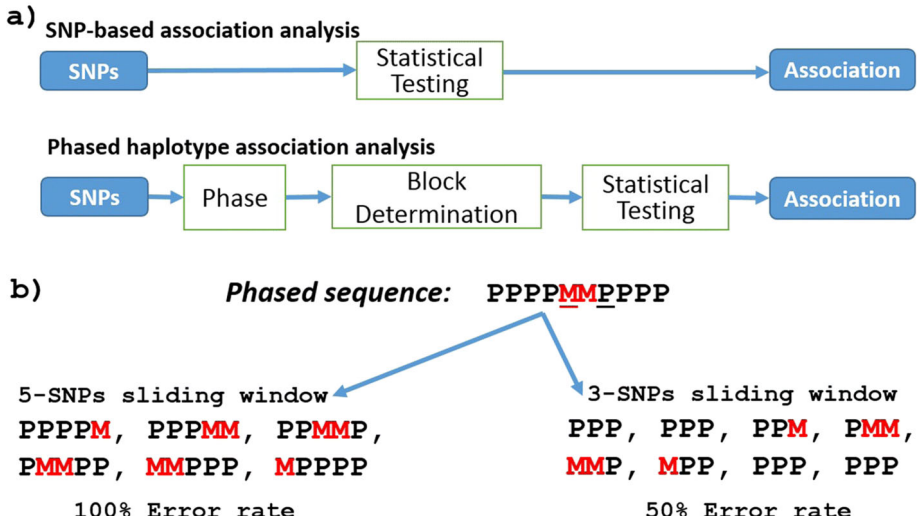



**Fig. 1** The relationship between phasing and block determination in haplotype association analysis: **a** Outline of the core steps within a standard SNP-based and a haplotype association analysis, highlighting the additional complexity of the haplotype-based approach. **b** the effect of block determination given a phased sequence. Here, we show a haplotype from 10 heterozygous SNPs. While all the alleles should all be allocated to the same allele, 2 have been allocated to the maternal v (marked as M in red) and the remaining to the paternal sequence (marked as P in black). As such, the phased sequence has 2 switches (phasing errors) at the loci 5 and 7 and hence has a switch error of 2/10 (20%). If we chose a 3-SNP sliding window with one SNP step, we find that 4/8 (50%) of the blocks contain errors. However, if a 5-SNP window was used, all blocks will contain errors. Using different window sizes affect the error rate within the blocks, yet switch error calculated at the whole sequence scale is constant 20%

three main steps: phasing, block determination and statistical analysis. A wide range of algorithms have been designed for phasing of the genome, with recent algorithms scaling to hundreds of thousands of individuals [11]. Block determination, overlapping or non-overlapping, typically uses a fixed-size sliding window [9] or is based on linkage disequilibrium (LD) so that each block contains alleles that are more likely to be inherited together [12, 13]. Finally, the phased blocks are assessed statistically to determine significant association with disease.

In haplotype association analysis, replicability of detected associations will be affected by the error rate within the phased haplotype blocks, which is dependent on both the error rate of the chosen phasing method and the approach chosen for block determination. An example of how block determination methods impact error rates is shown in Fig. 1b, where the phased sequence contains switch errors (i.e. it swaps between paternal and maternal allocations) at loci 5 and 7. However, the proportion of haplotypes blocks that contain phasing errors varies from 50% to 100% depending on whether a 3- or 5-SNP sliding window (shifted by one SNP at each step) approach is used. This simple example shows how both the initial phasing accuracy and the blocks determination method selected can impact the accuracy of haplotypes within a set of determined blocks.

Current haplotype association studies assume that haplotypes within the determined blocks (either LD or sliding window based) do not contain errors due to phasing [9, 12, 13]. This assumption may increase false positive associations [3, 14] given that phased haplotypes tend to be accurate only in short regions and this accuracy cannot be maintained for long regions [15, 16]. Some genomic regions contain more errors than others due to factors such as differences in linkage disequilibrium, recombination rate, and the density of SNPs [17]. The impact of this assumption on downstream analysis is unclear as the error rate of phased haplotype blocks has not been explored.

Existing evaluations of phasing tools have been conducted without considering the intended application of the phased haplotypes. Such studies evaluate phasing methods using metrics such as switch error, missing error, incorrect genotype percentage, and performance time [15, 16, 18, 19]. While these metrics are informative, they are typically reported as aggregates from across either a set of genomes or a single individual's entire genome. Such summary statistics do not necessarily reflect the quality of the phased haplotypes within specific regions or blocks, which is critical for downstream haplotype block analysis. Furthermore, these evaluations do not consider the downstream application of phased haplotypes and hence do not consider the joint impact of chosen phasing and block determination methods on phasing error rate of resulting haplotype blocks.

In this paper, we present the first evaluation of the behaviour of state-of-the-art phasing tools, as it relates to the direct use of phased haplotype blocks in

downstream association analysis. We evaluate seven well-known population-based haplotype estimation methods (fastPHASE, BEAGLE, IMPUTE, MaCH, SHAPEIT2, HAPI-UR, and EAGLE2) in addition to consensus haplotype estimators, which combine results from multiple phasing tools. We examine the interaction between phasing tool and two block determination approaches, based on sliding windows and LD thresholds, and their joint impact on error rates of derived haplotype blocks. We consider two different scenarios when phasing a particular region; either phasing the entire chromosome and then extracting the region, or extracting the region and then conducting phasing. Finally, the stability of the tools when applied to the same datasets several times is reported for all evaluation metrics.

## Results

### Different error locations obtained by different phasing tools

The switch errors observed across the six tools in chromosome 1 occurred at 12,145 different loci out of a possible 36,923 heterozygous SNPs. The distribution of error similarity by tool shown in Fig. 2a. The majority of these switches were unique to a single tool (∼56%), while only ∼5% of these loci were common between all tools. This large variation between tools in the sites of the switch errors implies that most tools are likely to result in different haplotypes being formed and hence may have a strong impact on haplotype analysis.

Examining the phasing of individual chromosomes highlights the variability of phasing tools. This variability is not reflected in overall summary metrics of phasing error. Figure 2b illustrates the different estimated haplotypes obtained by the six considered tools for a random individual within the same region. The example shows a contrast between the summary metrics and the resulting haplotypes. For example, EAGLE2 and BEAGLE both had a switch error of 5.4% (20 switches/372 heterozygous SNPs) in the same example, yet the estimations are different. Such examples motivate the development of metrics that may be more relevant to phased haplotype association analysis and that capture the error rate of the haplotype blocks used for downstream statistical analysis. Thus, we find that it is difficult to judge which estimation is better without considering the downstream application that will make use of the phase information.

### A consensus haplotype method improves phasing accuracy

The differences between the tools encouraged us to construct consensus haplotypes from the output of different tools. Due to its long runtime, MaCH was only examined

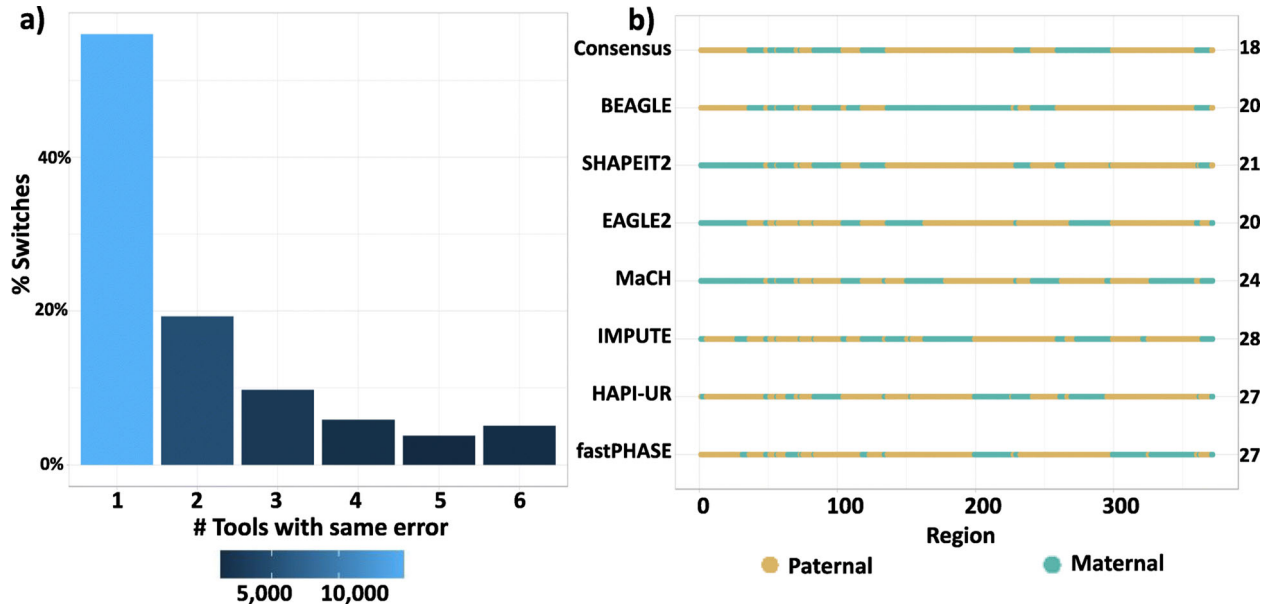



**Fig. 2** Similarity of switch location across phasing tools: **a** The frequency of shared switch error sites by different numbers of tools. The first bar represents the frequency of the error sites reported by one tool. The second bar represents the frequency of the error sites reported by two tools, and so on. **b** Estimated haplotypes for a random individual from seven different tools. The shown sequence corresponds to 372 heterozygous SNPs out of 1400 SNPs in the region Chr17:11045667-17395608. Each line represents a haplotype estimated by the corresponding tool, with the complementary phasing not shown (paternal regions become maternal and vice versa). Green represents the correctly phased haplotype runs that are identical to the maternal haplotype, while the brown runs are identical to the paternal ones. The numbers aligned to each tool name are the count of switches between the paternal and maternal copies

on chromosome 17 and thus was not considered in any consensus combinations.

The results in Table 1 shows that no single tool obtained the best results for the three chromosomes. The minimum switch error is obtained by SHAPEIT2 for chromosome 1, BEAGLE for chromosome 6, and EAGLE2 for chromosome 17. These three tools were found to be consistently always substantially better than the remaining four tools, with fastPHASE demonstrating the highest switch error.

Taking advantage of the independent phasing outputs between the individual tools, the best consensus, combining SHAPEIT2, EAGLE2 and BEAGLE, improves the accuracy for all datasets, with a 13% switch error improvement compared to the best single tool. However, if one of these tools is replaced with a less accurate tool (fastPHASE, IMPUTE and HAPI-UR), the consensus may provide worse results than one of the individual tools. Moreover, the addition of less accurate tools in consensus construction, as exemplified by the consensus using 5 tools, can also increase the overall error rate. Given the strong initial results for the consensus formed by SHAPEIT2, EAGLE2 and BEAGLE, we consider this approach in the rest of the tests reported in this study.

**Table 1** Switch error (%) obtained by the tools when applied on chromosomes 1, 6, and 17

| Approach | Tool | Chr1 | Chr6 | Chr17 |
|---|---|---|---|---|
| Individual | BEAGLE | 1.39 | **1.12** | 2.03 |
| | SHAPEIT2 | **1.33** | 1.20 | 2.06 |
| | EAGLE2 | 1.36 | 1.6 | **1.90** |
| | HAPI-UR | 2.14 | 1.81 | 3.00 |
| | IMPUTE | 2.83 | 2.46 | 4.30 |
| | fastPHASE | 4.10 | 3.43 | 5.32 |
| | MaCH | - | - | 4.10 |
| Consensus | SHAPEIT2, EAGLE2 and BEAGLE | **1.14** | **0.98** | **1.68** |
| | SHAPEIT2, EAGLE2, BEAGLE, IMPUTE and HAPI-UR | 1.2 | 1.04 | 1.76 |
| | SHAPEIT2, EAGLE2 and HAPI-UR | 1.24 | 1.08 | 1.79 |
| | SHAPEIT2, BEAGLE and fastPHASE | 1.37 | 1.14 | 2.06 |
| | SHAPEIT2, EAGLE2, IMPUTE, fastPHASE and HAPI-UR | 1.41 | 1.19 | 2.1 |
| | EAGLE2, IMPUTE and HAPI-UR | 1.48 | 1.27 | 2.16 |
| | SHAPEIT2, fastPHASE and HAPI-UR | 1.66 | 1.43 | 2.41 |
| | IMPUTE, fastPHASE and HAPI-UR | 2.16 | 1.82 | 3.19 |

Numbers in bold are the minimum error obtained according to each approach while underlined numbers are the minimum error obtained in all applied tests. MaCH was applied only on chromosome 17 due to the intensive performance time needed to phase other chromosomes. The tools are sorted according to the average switch error. The top section of the table lists the performance of the tools individually, while the bottom half lists the performance of the consensus estimators based on different combinations of phasing methods

## Accuracy of haplotype blocks varies according to the block determination method

### *An evaluation of haplotypes obtained by a sliding window*

We investigated the incorrect haplotype block percentage (IHBP, see Evaluation Criteria in Methods) obtained by a sliding window applied on chromosomes 1, 6 and 17. Figure 3 shows the strong impact of the window size on IHBP for all tools, which varied from an average IHBP of 2% to more than 50% when the window length increased from 5 SNPs to 100 SNPs. The haplotypes obtained by the consensus approach had the minimum IHBP for all window sizes, while EAGLE2, SHAPEIT2, and BEAGLE clearly outperform the remaining phasing approaches. The difference of the average IHBP (for all datasets and windows) between the consensus haplotype and the best single tool is almost 10 times higher than the difference between the best two single tools. While we do not attempt to define an optimal window width for all cases, longer haplotype windows are always more likely to contain a phasing error compared to shorter windows as shown in Fig. 3. Moreover, Fig. 3 indicates that a long sliding window approach, as used in previous work [9], may have a high error rate, when a comparable sample of unrelated individuals is used.

Since the error rate of phasing tools is likely to vary dramatically given differences in recombination rates, heterogeneity and SNP density, we have also considered the relationship between the phasing error rate and the number of incorrect blocks for the seven haplotype estimation tools. We sample regions of 1400 contiguous SNPs 50 times from chromosome 17, this time categorizing them based on the length of their correctly phased runs, i.e. the average number of contiguous SNPs that are correctly phased in respect to each other.

Figure 4 illustrates how, for a fixed window size, the number of incorrect haplotype blocks increases for more difficult to phase regions, where phasing difficulty is measured using the length of correctly phased runs. Taking EAGLE2 as an example, we see that in regions that have longer correctly phased runs (44 to 64 consecutive correctly phased SNPs), only 2% of haplotype blocks contain errors. In contrast, in more difficult to phase regions (24 to 30 consecutive correctly phased SNPs), a median of 3% of blocks contain errors. While these results indicate that error rates between different regions of the same genome can vary by over 50%, we note that similar observations can be seen between regions on different chromosomes.

In these scenarios, the consensus method had the minimum IHBP for different window sizes, and different

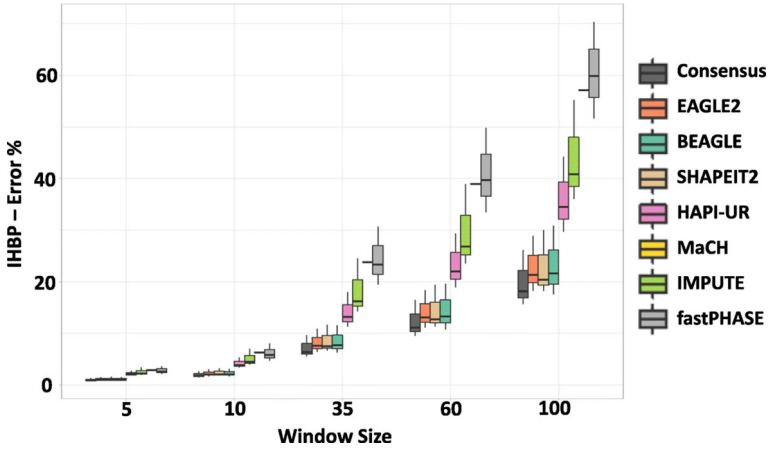



**Fig. 3** Impact of sliding window size on IHBP: Box plot summarizing the incorrect haplotype blocks percentage (IHBP) when applying different window sizes (5, 10, 35, 60, and 100 SNPs) to chromosomes 1, 6, and 17. The x-axis represents the window size, while y-axis represents shows IHBP. MaCH was only applied on chromosome 17 due to the extensive performance time required to phase chromosome 1, and 6, therefore its results are represented as a line

regions. Moreover, there is no result for the haplotype obtained by the consensus method when the length of the correctly phased runs is less than 30 SNPs. This observation demonstrates the efficacy of the consensus method as it was able to improve the accuracy and increase the length of the correctly phased runs to exceed this range. EAGLE2, BEAGLE, and SHAPEIT2 were the best tools individually.

### *An evaluation of haplotypes blocks obtained by ID*

The phasing error rate of blocks defined based on LD, measured again as IHBP across chromosome 1, 6, 17, is summarized for the different haplotype estimation tools in Table 2. Consistent with results in the previous sections, the consensus haplotype caller had the minimum IHBP while SHAPEIT2, BEAGLE and EAGLE2 showed substantially less error than the remaining individual tools. Around 50% of the blocks incorrectly phased by HAPI-UR are caused by incorrect imputation of at least one missing SNP located in the block.

In order to compare the error rate for sliding window and LD based block determination approaches, we used a sliding window with a width of 5 SNPs, equal to the mean of SNP count with blocks determined by LD. Both approaches were applied on chromosome 1, 6, and 17, and the incorrect haplotype block percentage was calculated for each tool and summarized in Fig. 5. We see that the error rate was halved in many cases when using an LD

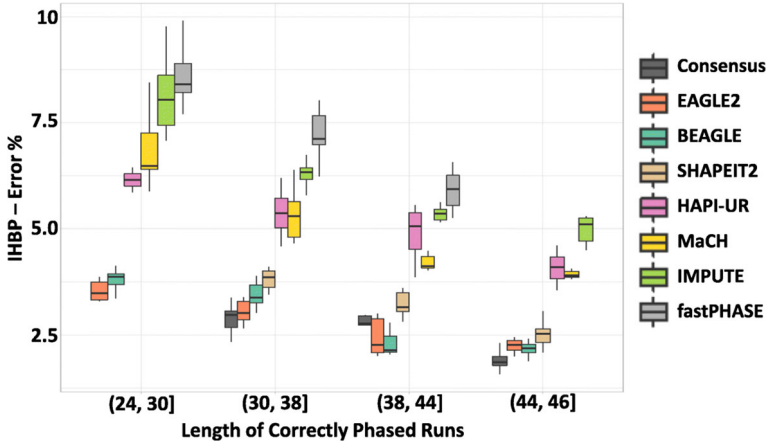



**Fig. 4** The relationship between the length of the correctly phased runs and IHBP: The incorrect haplotype blocks percentage (IHBP) was calculated for a fixed window size (10 SNPs) applied to 50 datasets from chromosome 17. The x-axis shows the correctly phased runs, binned according to the quartiles of the data. The y-axis shows IHBP. There is no bar for the consensus method in the range (24,30] as the length of the correctly phased runs always exceeds this range

**Table 2** Incorrect haplotype block percentage (%) obtained by the tools when applied on chromosomes 1, 6, and 17

| Tool | Chr1 | Chr6 | Chr17 |
|---|---|---|---|
| Consensus haplotype | **0.45** | **0.41** | **0.46** |
| BEAGLE | 0.46 | 0.43 | 0.49 |
| SHAPEIT2 | 0.47 | 0.42 | 0.51 |
| EAGLE2 | 0.48 | 0.44 | 0.48 |
| HAPI-UR | 1.26 | 1.27 | 1.25 |
| IMPUTE | 1.17 | 1.18 | 1.47 |
| fastPHASE | 0.54 | 0.5 | 0.58 |
| MaCH | - | - | 0.54 |

Numbers in bold are the minimum error obtained by any approach while underlined numbers are the minimum error for a single tool. MaCH was excluded from the tests applied on chromosome 1, and 6 due to extensive performance time

block approach (with a variable length from 2 SNPs to 34) compared to that of a sliding window. However, the number of blocks to be evaluated is greatly altered between the approaches, with a median number of blocks 6056 (interquartile range (IQR): 4390–6546) vs 31,729 (IQR: 22,269–34,327) for the LD-based and 5-SNP sliding window approach, respectively. These results highlight the trade-off between accuracy and comprehensiveness that to need be made when selecting a block determination methods to use in phased haplotype analysis.

### Impact of including surrounding regions on haplotype estimation

Haplotype studies interested in specific regions or genes [9, 20, 21], often phase only these particular regions rather than the whole chromosome. Therefore, we investigated the error rates of specific haplotype blocks when phasing either the entire genome or just the regions containing blocks of interest using SHAPEIT2, EAGLE2 and HAPI-UR.

Figure 6 shows that there are only small difference in accuracy regardless of whether phasing is performed on entire chromosomes or only selected regions. The error rate obtained by SHAPEIT2 and EAGLE2 was reduced for all metrics when phasing the whole genome followed by extraction of the regions of interest; however, the magnitude of improvement was small. When including surrounding regions, the median of sliding window-IHBP was reduced from 2.4 to 1.0% when applying SHAPEIT2 and from 1.94 to 1.73% when applying EAGLE2. HAPI-UR had different behaviour, achieving better results when phasing the short regions, increasing the IHBP from 2.96 to 3.27% when using a sliding window approach. However, given the poor performance of HAPI-UR on other evaluations in this work, it is unclear whether this result is likely to generalize to other tools. These results indicate that phasing the entire genome is likely to lead to improved results compared to phasing only specific regions, but the improvement may not warrant the additional computation time.

### Tool stability

We observed that the outputs of many of the phasing methods evaluated changed across multiple runs when the input remained the same, with this instability having potential to affect the replication of downstream analysis. While this observation was not made for EAGLE2, similar instability could be observed by permuting the order of individuals in the dataset being analysed. To explore this instability, we conducted a stability test for five tools.

Figure 7a illustrates the consistency of identifying a SNP location as a switch error across 15 runs. This distribution shows that the majority of the errors were either

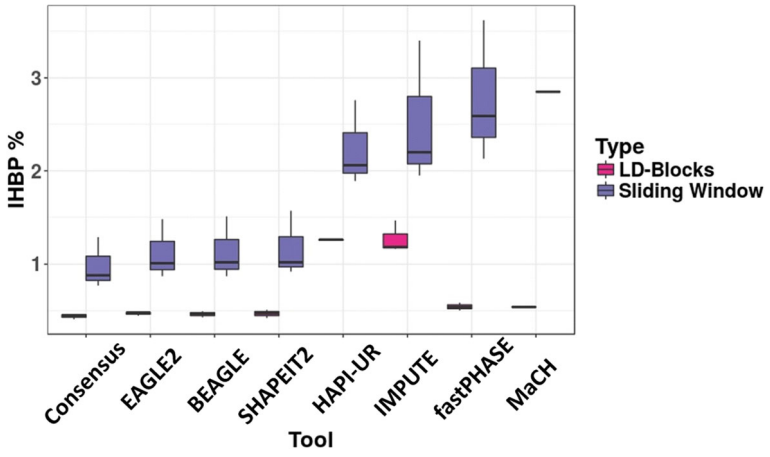



**Fig. 5** Comparison of haplotype block determination approaches. Boxplot showing the incorrect haplotype block percentage calculated for the blocks obtained by sliding window (5 SNPs width) or using an LD-based approach on the same dataset for each phasing method

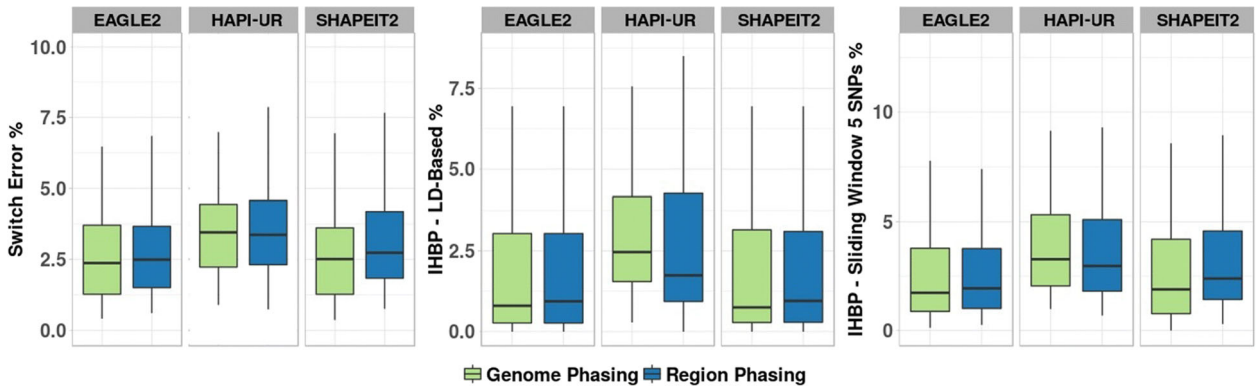



**Fig. 6** Evaluation Tests when including and excluding surrounding regions. Boxplots summarize switch error and IHBP with blocks determined using LD-based and sliding window (5 SNP) when phasing using SHAPEIT2, EAGLE2, or HAPI-UR is applied on the same regions in two different scenarios: one with including surrounding regions (green), and when only including target region (blue)

unique for each run, or were common across the 15 runs. The minimum error location variation was obtained by EAGLE2 followed closely by the consensus approach. The large proportion of errors unique to each run indicate that there is a substantial amount of variation between different runs of the same tool.

Figure 7b illustrates the differences between the estimations of SHAPEIT2, consistently one of the most accurate phasing tools, across the 15 iterations. This plot shows how the characteristics of the data influence the accuracy of the haplotype estimation. Here, we find that the estimation for the region from [350 to 500] was similar across most of the 15 iterations, while the results vary substantially for the region at [0 to 350]. Such variation is highly likely to affect downstream analysis and represents a significant issue for haplotype association analysis.

While Fig. 7 highlights the variability of the error locations, we can also examine how the instability of the phasing tools impacts the error rates of the haplotype blocks determined via sliding window and LD. Figure 8 illustrates the variance of errors obtained by each tool across multiple runs and highlights that there is substantial variability across different applications of the same tool. For instance, SHAPEIT2 shows switch error varying around 15% difference between the best and worst performing iteration of the tool. For all metrics, we again see the consensus approach obtains the best accuracy (in line with previous results). However, its variability

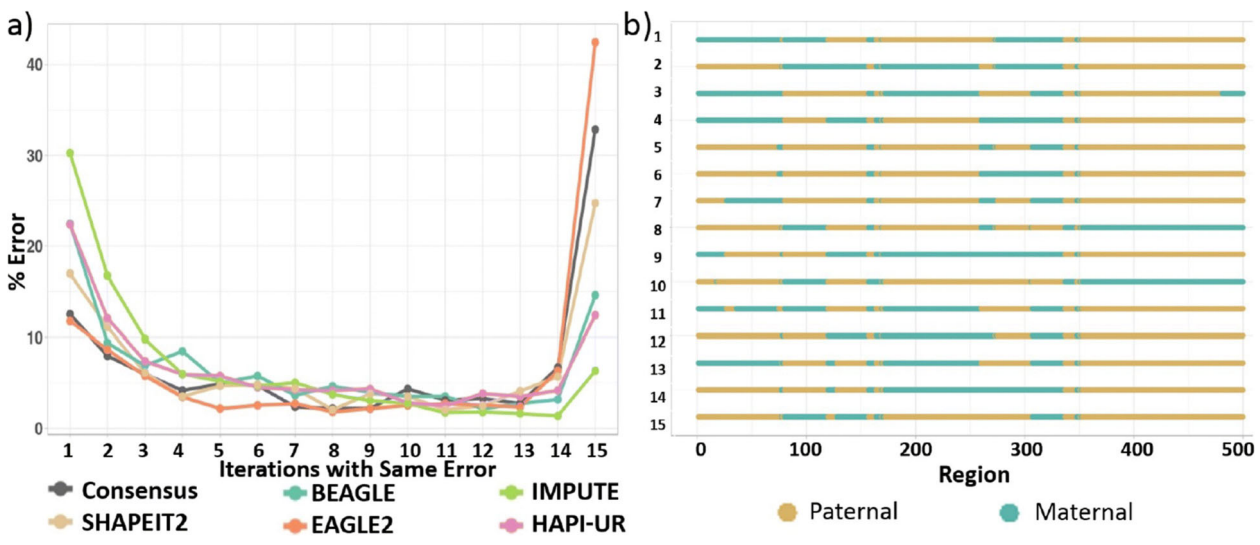



**Fig. 7** Stability of switch error. **a** The percentage of errors that are consistent across different runs of the same phasing method and same data. For example, the point (2, 9) in BEAGLE plot means that 9% of the errors reported by BEAGLE within 15 iterations occurred in the same location only in two iterations. **b** 15 different estimations for the same random individual (around 500 SNPs within the region chr1:212540742-224862598) obtained by SHAPEIT2 tool

is similar to BEAGLE and SHAPEIT2, as it is based on stochastic tools.

## Discussion

Improving the accuracy of phased haplotypes and understanding the impact of block determination methods on the accuracy of the haplotype blocks is critical to conducting accurate and replicable haplotype association analysis. This is the first study to evaluate phasing tools and their relationship with two standard methods for determining haplotype blocks with a focus on phasing accuracy at block scale. Our evaluations showed that three tools, SHAPEIT2, EAGLE2, and BEAGLE, consistently obtained the best results in all applied tests, with performance between these tools varying depending on the scenario and data under evaluation. This observation is in line with other studies that evaluated overall switch error [11, 19]. A consensus approach built from these three tools led to further improvements in switch error and incorrect haplotype block percentage. We demonstrated the trade-offs between phasing accuracy, block length and block count that are inherent between LD and sliding window based methods for block determination. Finally, we examined the stability of these tools and demonstrated that there is a large variability in outputs for most tools between runs.

### Ensemble approach improves haplotype estimation for all considered scenarios

While most previous evaluation of phasing tools focuses on overall switch error rates, our analysis compared the error profile from the different tools (Fig. 2). The comparison revealed that the switch error from different tools occurs in different locations, encouraging us to construct a consensus haplotype from the output of multiple phasing methods. A key outcome of this work is the robust phasing accuracy achieved by the consensus approach constructed from SHAPEIT2, EAGLE2 and BEAGLE across all metrics for all applied tests. This approach achieves a 13% improvement in switch error compared to the best individual tool and indicates a strong improvement compared to previous methodological phasing advances [22–24]. As expected, the performance of the consensus is influenced by the individual tools used in its construction, with less accurate tools lowering performance but remaining more accurate than any individual tool.

While the consensus of SHAPEIT2, BEAGLE and EAGLE2 gives the lowest error rates in all scenarios examined in this paper, the results in Table 1 show that there is no guarantee that such an ensemble approach will outperform its constituent tools individually, with the inclusion of fastPHASE in particular dragging down performance of any consensus it was included in. However, if the tools included in the ensemble are comparable in terms of accuracy, we observe that the resulting consensus is likely to lead to improvements in performance and robustness.

The downsides of the consensus caller approach are the increases in runtime, due to multiple passes over the same data. Hence, the tools may not scale effectively to large datasets. In this study, BEAGLE was the bottleneck of the consensus, taking one month to phase chromosome 1 for 12,008 individuals, while EAGLE2 and SHAPEIT2 accomplished the phasing task within a week, albeit using only a single thread of an AMD Opteron 6200. However, given that running multiple tools can be trivially parallelized and that haplotype phasing only needs to be run once for downstream analysis, the gain of the accuracy obtained from the consensus approach may justify the additional run time required.

### Trade-offs in block determination methods

The differences between the choice of LD and sliding window approaches for block determination in haplotype association analysis have long been recognized [17]. The former tends to produce fewer blocks whose length adapts to recombination properties of the genomic region under consideration, while the latter produces many, typically longer blocks with the same length across the genome. In this work, we have explored the impact that these trade-offs have on phasing error, focusing on IHBP to evaluate the proportion of blocks that are incorrectly phased. This metric more clearly elucidates the likely impact of phasing errors and block determination on the downstream application of statistical association analysis. When using a fixed size sliding window, our results showed a significant impact of the window size on the error rate of the haplotype blocks (Fig. 3), with error rates varying significantly depending on the region being phased (Fig. 4). The error rate variation, influenced by factors such as the recombination rate and the density of SNPs, demonstrates that a fixed size sliding window is likely to lead to many haplotypes that contain errors. This is highly likely to have a strong impact on the false negative and false positive rates of downstream analysis and is likely to place some limitations on the replicability of results.

In order to minimize the error rate within the haplotypes obtained by a sliding window, the width of the window should suit the length of the correctly phased runs. As there is no optimal way to determine the best window width, one solution may be an adaptive window size whereby the size of the window is adjusted according to the characteristics of the scanned region (e.g. is smaller if recombination rates are high or SNP density is low). Such an approach has been explored in the literature [25] and our results lend further support for such an approach. Alternatively, an LD-block approach may be used as this inherently adapts to the properties of the data, albeit with a number of settings that need to be defined by the user.

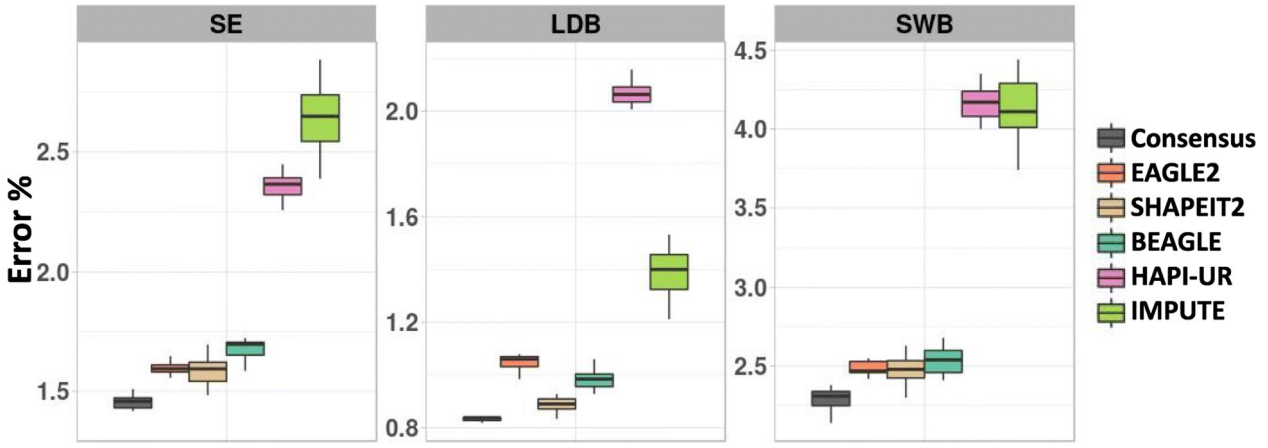



**Fig. 8** Impact of phasing method instability on error rates. Boxplot summarizing the variation of the obtained errors when applying the tools on the same dataset 15 times. The plots are, from left to right: switch error (SE), IHBP calculated for blocks determined according to the LD (LDB), and IHBP calculated for blocks determined by sliding window with a width of 10 SNPs (SWB)

We found that such an approach led to a strong reduction in switch error rates and the number of defined blocks that contain errors, at a cost of producing fewer, shorter block segments. This may be detrimental if the variation of a single region is needlessly broken into multiple independent blocks. We note it may be possible to use the consensus approach to determine block boundaries guided by whether individual methods are consistent at a given position or even to derive probabilistic block boundaries by treating the consensus caller as an ensemble predictor [26]. Combining this with probabilistic tests for haplotype association analysis [3, 14] may further reduce type 1 error rates in downstream analysis.

#### Tool stability

All algorithms for phase estimation considered in this work are based on Hidden Markov models which are often trained using a stochastic algorithm, which may explain why most phasing methods explored have different outputs across multiple runs when the input and parameters remains constant. Only EAGLE2 produces stable outputs by default. Although the variation in the error rate for most phasing methods was small across different iterations, the switches or the boundaries of correctly phased runs were affected significantly. The majority of the switches in the estimated haplotypes were either common across 15 iterations or unique to each iteration. We note that while some of the tools allow for a random seed to be passed in as a parameter, the default setting for the methods uses random initialization and that such seed parameters typically have no effect if the tools are run in a multi-threaded mode. Moreover, small changes in the input data, such as reordering samples, also appears to change the phasing output. As such, we believe that the instability observed in this study reflects the way that these phasing tools are being used in most practical applications.

The consensus haplotypes not only improved the accuracy but also had very stable results, rivalling those of EAGLE2. This instability of error location can be seen as similar to the difference in error location shown in Fig. 1b). As such, it may also be possible to exploit the instability of the tools to construct a consensus haplotype from the same tool applied several times on the same dataset. In this way it may be possible to yield the robust results from the consensus, while only running a single efficient tool.

An important question that arises here is whether the instability of the tools has a significant influence on detected results. Given that we observe relatively high error variation across the same data, one could imagine that the different tools may vary even more dramatically across different datasets. This is likely will not affect EAGLE2, the only stable tool, as its stability is based on the seeding conditions for a given set of input data and when that data changes, variation may occur. This variability in output is likely to be a key limitation on the replicability of haplotype association analysis. Future research could explore how the stochasticity of phasing algorithms could be exploited to reduce phasing error rates.

#### The impact of including surrounding regions on phasing

Haplotype studies interested in specific regions or blocks, such as those focusing on genes only [20, 21] or the replication of association analysis after finding genomic regions with significant associations [9], often phase only these particular regions rather than the whole genome. However, the choice to only focus on select genomic regions and phase their blocks, as opposed to phasing the genome and then extracting blocks, may affect

the quality of phasing. We have found that SHAPEIT2 and EAGLE2 perform better when including surrounding regions while HAPI-UR performed better when phasing only the region of interest. While EAGLE2 is the only stable tool according to the stability tests (using default parameters), it obtained different results for the same region when including or excluding surrounding regions. These results indicate that phasing the entire genome is likely to lead to improved results compared to phasing only specific regions. However, the improvement may not warrant the additional needed computation time.

### Limitations of the study

Our primary goal was to evaluate the interaction between the choice of phasing tool and the block determination algorithm and the impact these had on phasing error rates. In particular, the analysis conducted in this work did not explicitly explore the impact that these had on downstream haplotype analysis, given the large amount of variability that this could entail (different choices of statistical tests, different assumptions of genetic architecture, etc.). In order to reduce the number of analyses conducted, we have also limited the parameter options of both the phasing tools selected and the block determination algorithms to their default parameters unless otherwise explicitly stated. Finally, all error rates produced here were on a small subset of individuals for whom phase information was available (n=39). While this sample size is comparable with previous evaluations [11, 27], increasing the number of individuals for whom resolved phased information is available would help further refine these results and the differences between evaluated tools.

## Conclusions

This study reports on the interaction between the choice of phasing tool and the block determination algorithm, with critical implications for the application of phased haplotype blocks such as haplotype association analysis. We provide a comprehensive evaluation of seven different haplotype phasing tools (fastPHASE, BEAGLE, IMPUTE, MaCH, SHAPEIT2, HAPI-UR, and EAGLE2). We further introduce a consensus haplotype estimator based on combining output from multiple phasing tools, that achieved the lowest error rates across all scenarios considered.

The work provides guidance and evidence for the key constituent methods of haplotype analysis at block scale by showing the positive and negative consequences of each choice independently and when used together, as well as highlighting the possibility of tool instability. The insights provided by this work should inform future haplotype-based analyses as well as drive methodological research into phasing tools.

## Methods

### Datasets and preparation

Real haplotype data for a population is not readily available for comprehensive evaluation. One common approach to determine phasing data is to use data from trios to resolve child haplotypes based on the sequence of the parents [11, 16]. In this work, we make use of trios obtained from HapMap project [28] using Utah Residents (CEPH) with Northern and Western European Ancestry (CEU). This dataset contains 19 unrelated individuals and 39 complete families consisting of father, mother and child (117 individuals, of these 39 children were included while 78 parents were excluded for our study). Parent data were used to resolve child haplotypes but were otherwise excluded as our focus is on phasing unrelated individuals.

For phased children, 35% of SNPs were heterozygous, 0.3% were missing (unknown whether heterozygous or homozygous) and 0.08% were Mendelian errors (inconsistent alleles among trios). Using the parent information, we could resolve 80% of the child heterozygous SNPs, and 40% of the missing SNPs. Heterozygous SNPs for each child were resolved deterministically using the parent information via "phasing by transmission". A child's heterozygous and missing SNPs are resolved when at least one of its parents has a homozygous SNP in the same genomic locus. Haplotypes for the 39 children were extracted from chromosomes 1 (36,923 SNPs), 6 (31,727 SNPs), and 17 (12,807 SNPs). The restriction to three chromosomes was related to the high runtime needed by some of the tools under evaluation.

The genotype data for the 39 CEU children and 19 unrelated individuals were combined with 11,950 individuals from coeliac disease dataset (EGA accession: EGAS00000000057). While the additional genomic data does not have resolved phase information, the large number of samples has been shown to improve the accuracy of haplotype estimation [16] and enables us to emulate the performance of phasing tools in a real-life scenario. The coeliac disease dataset contains 11,950 individuals (cases and controls) and 528,969 SNPs (Illumina Hap 550). Details on the collection, and quality control procedure applied on the dataset is described previously [29].

Evaluation and error metrics calculation were computed only from the 39 children with partially known haplotypes, while the haplotype estimation tools were applied to the entire dataset of 12,008 individuals, in order to maximise phasing accuracy.

### Haplotype estimation methods

The seven haplotype estimation tools evaluated in this study are summarized in Table 3. These tools are all based on the probabilistic Hidden Markov Model (HMM) framework of Li and Stephens [30]. However, a direct

**Table 3** Population-based haplotype estimation tools used in this study

| Tool-Version | Year | Heuristic to reduce haplotype search space |
|---|---|---|
| fastPHASE - 1.4.8 | 2006 | Uses a haplotype-clustering model, where the set of all possible haplotypes are clustered into a small fixed number of "ancestral" haplotypes [31]. |
| BEAGLE - 4.1 | 2017 | Uses a haplotype-clustering model with a variable number of clusters, depending on the region under consideration [24]. |
| IMPUTE - 2.3.2 | 2009 | Subsamples possible haplotypes that are similar to those of the currently estimated haplotype of an individual [32]. |
| MaCH - 1.0.18.c | 2010 | Subsamples possible haplotypes from the set of all possibilities randomly at each iteration [33]. |
| SHAPEIT2 - v2.r837 | 2012 | Breaks the chromosome into small windows of a few SNPs, estimates the phase of each window using a method similar to IMPUTEv2 and then estimates transitions between windows [23]. |
| HAPI-UR - v1.01 | 2012 | Breaks chromosome into small windows, that are initially very short but iteratively grow to a user defined size, enabling modelling of longer segments at once [34]. |
| EAGLE2 - v2.3.5 | 2016 | When no reference panel is provided (the scenario in this study), EAGLE2 applies long range phasing (EAGLE1 [27]) then efficiently represents all haplotypes such that beam search can be applied to evaluate only the most promising phase paths [22]. |

application of the Li and Stephens approach scales linearly with the number of individuals and the number of loci and quadratically with the number of possible haplotypes, limiting its direct applicability to large datasets. Given this, phasing tools have introduced heuristics to reduce the computational cost of phasing while trying to limit the impact on phasing accuracy. A brief summary of the heuristic used by each tool is given in Table 3. All the tools were applied using their default parameters. Genetic maps, which contain information about recombination rates across the genome, were used as an additional parameter to most tools (all except fastPHASE and MaCH) with EAGLE2 taking a specific format[1] while all other tools used a PLINK format[2]. All genetic maps were made with respect to the GRCh36 reference genome (genome build hg18).

### Consensus haplotype construction

We constructed a consensus haplotype of the estimations obtained by several combinations of three and five tools. The consensus haplotype is assembled as follows:

1. The haplotypes of the tools are aligned to each other (one copy of the estimated haplotype pair from each tool). The haplotypes should have the same allele at the first heterozygous SNP.
2. Scan the haplotypes SNP by SNP, and for each SNP, vote for the alleles and choose the final one (for the consensus estimator) according to the majority.
3. If the allele at the scanned SNP for a tool doesn't agree with the final allele (based on the majority of tools), switch the copies of the chromosome pair for the remaining SNPs for this specific tool.

### Evaluation criteria

The standard metric to assess the accuracy of haplotype estimation is switch error, the number of switches in the estimated haplotype divided by all possible switches (heterozygous SNPs count) [15, 16]. Most studies measure switch error as it reflects the accuracy with respect to the neighbouring SNPs [24]. Switch error is formally defined as:

$$SE = \frac{\sum_{i=1}^{n} \frac{switches}{H-1}}{n} \quad (1)$$

where *switches* is the number of the incorrectly phased heterozygous SNPs in comparison to their predecessor SNPs (SNPs in the previous genomic position), $H$ is individual's heterozygous SNPs count, and $n$ is the individuals count in the dataset.

In this study, we focus on the accuracy of phased haplotypes with respect to the block determination approach. Therefore, we introduce a new accuracy metric termed Incorrect Haplotype Block Percentage (IHBP) in order to calculate the error rate within blocks determined either by sliding window, LD between SNPs or any other approach. The formula for IHBP is defined as:

$$IHBP = \frac{IB}{B} \quad (2)$$

where $IB$ is the count of the incorrectly phased haplotype blocks and $B$ is the count of all haplotype blocks. All unambiguous blocks, i.e. blocks that contain only homozygous or one heterozygous SNPs, were also excluded. A haplotype pair within a block is considered correctly phased if all heterozygous and missing SNPs are phased correctly with respect to each other. In other words, one copy of the pair is identical to the paternal haplotype, and the second copy is identical to the maternal haplotype within the same block (though we do not explicitly identify the origin of each phased sequence).

When calculating error metrics, all unresolved SNPs (child's heterozygous SNPs when both parents have heterozygous SNPs in the same loci) and Mendelian errors were excluded from the evaluation. Missing SNPs that were resolved using family information were included in

[1] https://data.broadinstitute.org/alkesgroup/Eagle/downloads/
[2] http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/

the evaluation as all haplotype estimation tools used in this study implicitly impute missing SNPs.

An alternative metric of phase accuracy is length of correctly phased runs, defined as

$$LCPR = \frac{\sum_{i=1}^{n} LR_i}{n} \quad (3)$$

where $LR_i$ is the length of the correctly phased run $i$, and $n$ is the number of the correctly phased runs within a region. A correctly phased run refers to contiguous sets of correctly phased heterozygous SNPs with respect to each other. The length of the correctly phased runs within each region was calculated as the mean of the number of heterozygous SNPs within each run.

#### Haplotype blocks determination via sliding window

The sliding window scans the whole chromosome in this work shifting one SNP step at a time, treating each window as a haplotype block. The possible haplotype blocks within a sequence of $m$ SNPs are $m-w+1$ blocks, where $w$ is the window size. Phasing accuracy of blocks determined by a sliding window with five different random sizes (5, 10, 35, 60, and 100 SNPs) and one SNP step was investigated in this study with respect to chromosomes 1, 6 and 17.

#### Linkage disequilibrium (LD) based haplotype blocks determination

In this study, LD-based blocks were determined which implements the confidence interval (CI) algorithm [35] as implemented by PLINK (v1.90b4.4) [36]. The algorithm requires a number of heuristic thresholds to be set and, in line with previous haplotype studies [12], we have made use of the following default parameters: SNP pairs are considered if they are within 200 kilobases (kb) of each other. SNPs with minor allele frequency (MAF) $< 0.05$ were excluded. SNP pairs are considered highly correlated (belong to the same block) if the bottom of the 90% $D'$ confidence interval is $>0.70$, and the top of the confidence interval is $>0.98$.

To examine the error rate of the resulting blocks, we consider the IHBP from ∼4800 LD blocks derived from chromosome 1, 6, and 17 for the 39 CEU children. This analysis excludes haplotype blocks which contained fewer than 2 heterozygous SNPs (i.e. which blocks which needed no phasing). The average SNP count within these blocks is 4.3 SNPs (minimum 2 SNPs and maximum 34 SNPs). The average length of these blocks is 17.5 Kb (ranging from 0.005–200kB).

#### Analysis of including surrounding regions on phasing particular regions

Two scenarios were applied for this investigation:

1. Phasing then block determination: Phasing methods were applied on the whole chromosome, then the estimated haplotypes were extracted to obtain the targeted region.
2. Block determination then phasing: Phasing methods were applied on specific regions without including any other neighbouring parts of the chromosome.

60 datasets were constructed by randomly selecting 150 or 250 contiguous SNPs from chromosome 1, 6, and 17 from 6000 individuals (including all 39 CEU children). Short regions were chosen as excluding or including surrounding regions will affect phasing accuracy in the boundary of the region of interest, and also for execution time issues as this evaluation was applied to 60 different regions. Only the fastest tools (SHAPEIT2, EAGLE2, and HAPI-UR) were used for this test. We compared the error rate within the results in both scenarios for all accuracy metrics. Both Switch Error and IHBP were calculated for each scenario, with IHBP calculated for blocks determined by 5-SNP sliding windows and based on LD.

#### Stability testing

The tools were applied to a randomly selected region of 2000 heterozygous SNPs (chr1:212540742-224862598) for 4000 individuals. The size of the region was limited to this size due to computational constraints, based on the time needed to execute 5 tools 15 times. MaCH and fastPHASE were not used in this comparison due to excessively slow performance. EAGLE2 was stable when using its default parameters. Therefore, we created 15 different datasets of the same region but with shuffling the individuals randomly for each of them. Shuffling the individuals made EAGLE2 unstable and allowed the assessment of its behaviour. Three error metrics were calculated: switch error and IHBP based on either LD-based or 10-SNP sliding windows.

#### Abbreviations

CEPH: Centre d'Etude du Polymorphism Humain families reference panel; CEU: Utah Residents from the CEPH collection with Northern and Western European Ancestry; HAPI-UR: HAPlotype Inference for UnRelated samples; HMM: Hidden Markov Model; IHBP: Incorrect haplotype block percentage; IQR: Interquartile range; LCPR: Length of correctly phased runs; LD: Linkage disequilibrium; LDB: IHBP calculated for blocks determined according to the LD; MAF: Minor allele frequency; SE: Switch error; SHAPEIT: Segmented HAPlotype Estimation and Imputation Tool; SNP: Single-nucleotide polymorphism; SWB: IHBP calculated for blocks determined by sliding window

#### Acknowledgements

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

#### Authors' contributions

ZB: Prepared data and code, performed the study, generated all figures, and wrote the first version of the manuscript. ZB, JZ, KV, and BG: conceived and designed the study, analysed the results. ZB, KV, and BG: contributed to

revising the manuscript. AK: Contributed to the background of the study. All authors reviewed, revised and wrote feedback for the manuscript. All authors read and approved the final manuscript.

**Funding**
ZB is supported by Melbourne Research Scholarship (ref: 103500) provided by the University of Melbourne. The University did not play any direct role in the design of the study or any of collection, analysis, interpretation of data or writing the manuscript, outside of the contributions of the authors that are University staff (KV, JZ, AK) as outlined in the previous section.

**Availability of data and materials**
The datasets used and analysed during the current study are:

1. Coeliac disease dataset is available from European Genome-phenome Archive. Link: https://www.ebi.ac.uk/ega/studies/EGAS00000000057.
2. Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) genotypes are available publicly from HAPMAP project. Link: ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
KV serves as a Section Editor for the *Knowledge-based analysis* section of *BMC Bioinformatics*. No other competing interests are declared.

**Author details**
[1]School of Computing & Information Systems, University of Melbourne, 3010 Parkville, Australia. [2]Centre for Neural Engineering, University of Melbourne, 3053 Carlton, Australia. [3]Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland. [4]Centre for Epidemiology and Biostatistics, The University of Melbourne, 3010 Parkville, Australia. [5]IBM Australia - Research, 3006 Southgate, Australia.

Received: 25 February 2019 Accepted: 10 September 2019
Published online: 30 October 2019

**References**

1. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic Acids Res. 2016;45(D1):896–901.
2. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. Nat Rev Genet. 2011;12(3):215.
3. Tregouet D-A, Garelle V. A new java interface implementation of thesias: testing haplotype effects in association studies. Bioinformatics. 2007;23(8):1038–9.
4. Garnier S, Truong V, Brocheton J, Zeller T, Rovital M, Wild PS, Ziegler A, Munzel T, Tiret L, Blankenberg S, et al. Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. PLoS Genet. 2013;9(1):1003240.
5. Ying D, Li M, Sham PC, Li M. A powerful approach reveals numerous expression quantitative trait haplotypes in multiple tissues. Bioinformatics. 2018;1:6.
6. Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, Akan P, Stupka E, Down TA, Prokopenko I, et al. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the fto type 2 diabetes and obesity susceptibility locus. PloS ONE. 2010;5(11): 14040.
7. Brown R, Kichaev G, Mancuso N, Boocock J, Pasaniuc B. Enhanced methods to detect haplotypic effects on gene expression. Bioinformatics. 2017;33(15):2307–13.
8. Zakharov S, Wong TY, Aung T, Vithana EN, Khor CC, Salim A, Thalamuthu A. Combined genotype and haplotype tests for region-based association studies. BMC Genomics. 2013;14(1):569.
9. Howard DM, Hall LS, Hafferty JD, Zeng Y, Adams MJ, Clarke T-K, Porteous DJ, Nagy R, Hayward C, Smith BH, et al. Genome-wide haplotype-based association analysis of major depressive disorder in generation scotland and uk biobank. Transl Psychiatry. 2017;7(11):1263.
10. Pei X, Liu L, Cai J, Wei W, Shen Y, Wang Y, Chen Y, Sun P, Imam MU, Ping Z, et al. Haplotype-based interaction of the ppargc1a and ucp1 genes is associated with impaired fasting glucose or type 2 diabetes mellitus. Medicine. 2017;96(23):e6941.
11. O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, Zagury J-F, Delaneau O, Marchini J. Haplotype estimation for biobank-scale data sets. Nat Genet. 2016;48(7):817.
12. Wu Y, Fan H, Wang Y, Zhang L, Gao X, Chen Y, Li J, Ren H, Gao H. Genome-wide association studies using haplotypes and individual snps in simmental cattle. PLoS ONE. 2014;9(10):109330.
13. Shang Z, Lv H, Zhang M, Duan L, Wang S, Li J, Liu G, Ruijie Z, Jiang Y. Genome-wide haplotype association study identify tnfrsf1a, casp7, lrp1b, cdh1 and tg genes associated with alzheimer's disease in caribbean hispanic individuals. Oncotarget. 2015;6(40):42504.
14. Curtis D, Sham PC. Estimated haplotype counts from case-control samples cannot be treated as observed counts. Am J Hum Genet. 2006;78(4):729–31.
15. Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, et al. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet. 2006;78(3):437–50.
16. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12(10):703–14.
17. Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol Off Publ Int Genet Epidemiol Soc. 2007;31(5):365–75.
18. Miar Y, Sargolzaei M, Schenkel FS. A comparison of different algorithms for phasing haplotypes using holstein cattle genotypes and pedigree data. J Dairy Sci. 2017;100(4):2837–49.
19. Herzig AF, Nutile T, Babron M-C, Ciullo M, Bellenguez C, Leutenegger A-L. Strategies for phasing and imputation in a population isolate. Genet Epidemiol. 2018;42(2):201–13. Wiley Online Library.
20. Tello-Ruiz MK, Curley C, DelMonte T, Giallourakis C, Kirby A, Miller K, Wild G, Cohen A, Langelier D, Latiano A, et al. Haplotype-based association analysis of 56 functional candidate genes in the ibd6 locus on chromosome 19. Eur J Hum Genet. 2006;14(6):780.
21. Barendse W. Haplotype analysis improved evidence for candidate genes for intramuscular fat percentage from a genome wide association study of cattle. PloS ONE. 2011;6(12):29601.
22. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. Reference-based phasing using the haplotype reference consortium panel. Nat Genet. 2016;48(11):1443.
23. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9(2):179–81.
24. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.
25. Guo Y, Li J, Bonham AJ, Wang Y, Deng H. Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. Eur J Hum Genet. 2009;17(6):785.
26. Zhong W, Kwok JT. Accurate probability calibration for multiple classifiers. In: Twenty-Third International Joint Conference on Artificial Intelligence. Beijing: AAAI Press; 2013. p. 1939–45.
27. Loh P-R, Palamara PF, Price AL. Fast and accurate long-range phasing in a uk biobank cohort. Nat Genet. 2016;48(7):811–6.
28. Consortium IH, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52.
29. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Ádány R, Aromaa A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010;42(4):295.
30. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213–33.
31. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78(4):629–44.
32. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):1000529.

33. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34.
34. Williams AL, Patterson N, Glessner J, Hakonarson H, Reich D. Phasing of many thousands of genotyped samples. Am J Hum Genet. 2012;91(2): 238–51.
35. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. The structure of haplotype blocks in the human genome. Science. 2002;296(5576):2225–9.
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

## Publisher's Note

## 3.8 Additional information

In this part, we present some evaluation experiments that have not been reported in the published manuscript. In addition to the real data, the experiments involve genotypes data generated through simulations and random mating of chromosome X.

### 3.8.1 Additional data for phasing evaluation

In addition to the real datasets used for the evaluation in the manuscript above, we also evaluated haplotype phasing tools using genotype data simulated and formed through random mating of chromosome X.

1. **Simulated datasets:** Cosi2 tool (Shlyakhter et al., 2014) was used to simulate real haplotypes for four different populations using "best-fit" parameters provided by the tool and described in a publication from Schaffner et al. (Schaffner et al., 2005). SNPs with Minor Allele Frequency (MAF) $< 0.05$ were eliminated and the remaining were thinned to obtain 1 SNP per 4 kbp resolution. We randomly mate a pair of the simulated haplotypes to construct the genotype sequence that will be passed to the tools as an input. 3% of the SNPs were arbitrarily masked to simulate the missing SNPs. 12 simulated datasets were prepared for European, African, African-American, and Asian ancestries. Details about these datasets are described in the Supplementary materials for chapter 3 Table A.1.

2. **Sex chromosomes datasets:** As described in the background, section: 2.5.1 Haplotype data for evaluation, the haplotypes in males' sex chromosomes can be used as a source of real haplotypes. We used the European coeliac disease dataset accessed from `https://www.ebi.ac.uk/ega/studies/EGAS00000000057` to create these datasets. We filtered chromosome X for all male individuals, then we removed all individuals who have more than 10 heterozygous SNPs (most likely some errors introduced in the data, therefore, we considered them as missing SNPs when generating the genotypes). The sequences that passed the filtration are randomly mated in pairs to generate genotype sequences to form 4 different datasets with different length and population size. Details about these datasets are described in the Supplementary materials for chapter 3 Table A.2.

### 3.8.2 Evaluation criteria

We calculate switch error (SE), incorrect genotype percentage (IGP), and missing errors (ME) as described in the background of this thesis (section: 2.5.2 Evaluation criteria). Additionally, we also report **Linkage Disequilibrium Error (LDE)**, that refers to the quality of phasing with respect the linkage disequilibrium between SNPs. The main advantage of this measurement is estimating the correctness of phasing a pair of SNPs with respect to the linkage disequilibrium between them. We calculate LDE as the percentage of the incorrectly phased SNPs pairwise in different linkage disequilibrium levels as follow:

$$LDE_k = \frac{IP_k}{P_k} \tag{3.1}$$

Where $IP_k$: The number of the incorrectly phased pairs with a linkage disequilibrium belongs to the range $_k$, $P_k$: pairs with the linkage disequilibrium belong to the range $_k$. LD was estimated by calculating $r^2$ using PLINK software.

Only the pairs of heterozygous and missing SNPs are considered in this evaluation when they are less than 1 mbp apart. A pair is considered correctly phased if the haplotypes of both SNPs match the paternal and the maternal ones. When a pair contains a missing SNP, the missing SNP should be correctly imputed and phased with respect to the pairing SNP.

### 3.8.3 Evaluation results

All the tools performed very well when applied on the simulated data (shown in Tables 3.4, 3.5, 3.6, and 3.7), then less accurate on the random mating of chromosome X (shown in Table 3.8), and the worst results obtained when dealing with real data (shown in Table 3.9). This observation encouraged us to limit the results in the published manuscript to the application pf phasing to real data.

The consensus approach of SHAPEIT2, BEAGLE, and EAGLE2 obtained higher accuracy than any of the tools used in its construction in all tests when considering switch error.

The comparison of error metrics shows the limitation of IGP. IGP was reported to be high especially with long regions which does not mean having low-quality phasing. As

**Table 3.4: Phasing evaluation for simulated dataset for European population according to the percentage of SE, IGP and ME.**. Datasets in the first row are described in the Supplementary materials for chapter 3 Table A.1.

| Data / Tool | SD1 | | | SD2 | | | SD3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** |
| **Consensus** | 0.146 | 0.830 | 0.809 | 0.058 | 0.566 | 0.363 | 0.000 | 0.031 | 0.001 |
| **SHAPEIT2** | 0.178 | 1.390 | 1.033 | 0.110 | 0.731 | 0.712 | 0.010 | 0.336 | 0.022 |
| **BEAGLE** | 0.308 | 2.837 | 1.272 | 0.179 | 1.268 | 0.612 | 0.004 | 0.006 | 0.009 |
| **EAGLE2** | 0.226 | 1.425 | 1.081 | 0.148 | 0.958 | 0.668 | 0.005 | 0.463 | 0.089 |
| **HAPI-UR** | 0.853 | 7.179 | 38.653 | 0.592 | 4.400 | 35.091 | 0.342 | 12.620 | 49.01 |
| **IMPUTE** | 0.567 | 4.210 | 0.931 | 0.204 | 1.719 | 0.543 | 0.008 | 0.048 | 0.005 |
| **fastPHASE** | 0.624 | 5.218 | 1.247 | 0.561 | 6.457 | 0.980 | 0.437 | 19.405 | 0.951 |
| **MaCH** | 0.139 | 1.260 | 0.756 | 0.170 | 1.077 | 0.651 | 0.054 | 2.927 | 0.393 |

**Table 3.5: Phasing evaluation for simulated dataset for African population according to the percentage of SE, IGP and ME.**. Datasets in the first row are described in the Supplementary materials for chapter 3 Table A.1.

| Data / Tool | SD3 | | | SD4 | | | SD5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** |
| **Consensus** | 0.215 | 1.552 | 0.910 | 0.055 | 0.454 | 0.299 | 0.000 | 0.000 | 0.002 |
| **SHAPEIT2** | 0.268 | 1.514 | 1.182 | 0.088 | 0.580 | 0.533 | 0.002 | 0.093 | 0.004 |
| **BEAGLE** | 0.611 | 3.713 | 0.929 | 0.178 | 1.335 | 0.527 | 0.005 | 0.032 | 0.018 |
| **EAGLE2** | 0.564 | 2.716 | 2.321 | 0.094 | 0.969 | 0.599 | 0.000 | 0.012 | 0.036 |
| **HAPI-UR** | 1.703 | 10.410 | 38.706 | 0.828 | 5.396 | 35.971 | 0.235 | 6.285 | 48.22 |
| **IMPUTE** | 0.599 | 4.838 | 0.606 | 0.276 | 1.592 | 0.333 | 0.008 | 0.035 | 0.006 |
| **fastPHASE** | 1.011 | 9.920 | 2.428 | 0.730 | 7.491 | 1.451 | 0.725 | 21.840 | 1.754 |
| **MaCH** | 0.251 | 1.826 | 0.714 | 0.084 | 0.736 | 0.651 | 0.061 | 2.551 | 0.454 |

**Table 3.6: Phasing evaluation for simulated dataset for African-American population according to the percentage of SE, IGP and ME.**. Datasets in the first row are described in the Supplementary materials for chapter 3 Table A.1.

| Data / Tool | SD6 | | | SD7 | | | SD8 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** |
| **Consensus** | 0.187 | 0.817 | 0.746 | 0.052 | 0.429 | 0.295 | 0.000 | 0.000 | 0.002 |
| **SHAPEIT2** | 0.220 | 1.186 | 0.910 | 0.077 | 0.314 | 0.620 | 0.002 | 0.078 | 0.007 |
| **BEAGLE** | 0.689 | 3.734 | 1.214 | 0.148 | 1.302 | 0.530 | 0.006 | 0.010 | 0.007 |
| **EAGLE2** | 0.301 | 1.583 | 1.482 | 0.116 | 0.622 | 0.811 | 0.000 | 0.000 | 0.019 |
| **HAPI-UR** | 2.160 | 11.714 | 40.662 | 1.015 | 5.185 | 36.778 | 0.175 | 5.530 | 46.94 |
| **IMPUTE** | 0.475 | 3.806 | 0.678 | 0.191 | 1.453 | 0.380 | 0.003 | 0.029 | 0.005 |
| **fastPHASE** | 1.331 | 10.913 | 2.659 | 0.936 | 8.819 | 2.475 | 0.640 | 20.959 | 1.739 |
| **MaCH** | 0.214 | 1.603 | 0.850 | 0.086 | 0.970 | 0.518 | 0.040 | 1.762 | 0.425 |

a switch error in the centre of a region leads to IGP equal to ∼50%. Therefore, switch error is considered a better metric to reflect phasing quality in the majority of studies.

**Table 3.7: Phasing evaluation for simulated dataset for Asian population according to the percentage of SE, IGP and ME.**. Datasets in the first row are described in the Supplementary materials for chapter 3 Table A.1.

| Data / Tool | SD10 | | | SD11 | | | SD12 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** |
| **Consensus** | 0.131 | 0.784 | 0.740 | 0.066 | 0.796 | 0.284 | 0.000 | 0.017 | 0.003 |
| **SHAPEIT2** | 0.202 | 1.595 | 0.938 | 0.108 | 0.778 | 0.614 | 0.006 | 0.273 | 0.010 |
| **BEAGLE** | 0.270 | 2.142 | 0.674 | 0.143 | 1.102 | 0.472 | 0.004 | 0.047 | 0.015 |
| **EAGLE2** | 0.234 | 1.665 | 2.246 | 0.193 | 2.012 | 0.754 | 0.004 | 0.365 | 0.067 |
| **HAPI-UR** | 1.141 | 7.371 | 34.246 | 0.971 | 6.518 | 38.996 | 0.438 | 14.683 | 48.57 |
| **IMPUTE** | 0.558 | 3.789 | 0.782 | 0.244 | 1.899 | 0.536 | 0.002 | 0.028 | 0.005 |
| **fastPHASE** | 0.699 | 7.089 | 1.825 | 0.711 | 6.512 | 1.299 | 0.520 | 21.085 | 0.982 |
| **MaCH** | 0.098 | 0.534 | 0.616 | 0.170 | 1.124 | 0.587 | 0.067 | 3.050 | 0.373 |

**Table 3.8: Phasing evaluation for simulated dataset for dataset generated from males' chromosome X according to the percentage of SE, IGP and ME.**. These datasets do not contain missing SNPs with known real haplotypes to calculate missing error. Datasets in the first row are described in the Supplementary materials for chapter 3 Table A.2.

| Data / Tool | XD1 | | XD2 | | XD3 | | XD4 | |
|---|---|---|---|---|---|---|---|---|
| | **SE** | **IGP** | **SE** | **IGP** | **SE** | **IGP** | **SE** | **IGP** |
| **Consensus** | 3.215 | 16.852 | 2.360 | 15.763 | 0.167 | 4.235 | 0.002 | 0.595 |
| **SHAPEIT2** | 3.216 | 16.826 | 2.408 | 13.784 | 0.167 | 2.394 | 0.043 | 12.747 |
| **BEAGLE** | 4.115 | 21.184 | 3.246 | 20.348 | 0.720 | 14.030 | 0.022 | 0.507 |
| **EAGLE2** | 3.964 | 18.954 | 3.207 | 17.301 | 0.278 | 5.586 | 0.009 | 3.322 |
| **HAPI-UR** | 5.704 | 23.282 | 4.054 | 22.206 | 1.602 | 24.595 | 0.135 | 21.265 |
| **IMPUTE** | 4.526 | 19.785 | 3.630 | 19.411 | 0.872 | 28.354 | 0.401 | 41.855 |
| **fastPHASE** | 4.404 | 22.382 | 4.524 | 26.640 | 5.164 | 41.549 | 4.168 | 46.348 |
| **MaCH** | 3.418 | 17.754 | 3.479 | 22.210 | 4.053 | 39.153 | 3.042 | 45.192 |

With respect to ME, the consensus does not obtain the minimal error rate in all tests. When imputation is done, there can be three possibilities (0, 1, and 2) for each SNP without considering allele allocation. Having more than two options will affect the performance of majority voting of three tools as ties can happen. This can be an explanation of this case. HAPI-UR obtained very high missing error reaching 50%. The authors of HAPI-UR described in their publication how HAPI-UR deals with missing data but they did not report any results related to the missing error in their experiments.

The impact of LD on phasing accuracy is illustrated in Figure 3.9. The higher the LD the better the phasing, which is expected as most of the phasing tools rely on the observation that haplotypes are similar within short regions of highly correlated

**Table 3.9: Phasing evaluation for real dataset according to the percentage of SE, IGP and ME.**. Datasets used here are the same ones in the published manuscript, however, here we report IGP and ME that are not used in the manuscript. Details about the datasets are in the manuscript and in the Supplementary materials for chapter 3 Table A.3.

| Tool \ Data | Chr1 | | | Chr6 | | | Chr17 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** | **SE** | **IGP** | **ME** |
| **Consensus** | 1.140 | 45.04 | 2.855 | 1.676 | 42.75 | 1.261 | 0.981 | 41.63 | 3.195 |
| **SHAPEIT2** | 1.332 | 43.84 | 2.971 | 2.060 | 41.67 | 1.986 | 1.199 | 42.29 | 2.768 |
| **BEAGLE** | 1.389 | 45.07 | 2.810 | 2.026 | 42.21 | 1.437 | 1.120 | 41.40 | 3.561 |
| **EAGLE2** | 1.364 | 44.07 | 3.232 | 1.904 | 43.82 | 2.855 | 1.164 | 43.12 | 4.012 |
| **HAPI-UR** | 2.144 | 45.77 | 71.956 | 3.002 | 44.77 | 72.108 | 1.808 | 44.34 | 62.67 |
| **IMPUTE** | 2.833 | 46.56 | 3.861 | 4.302 | 44.79 | 2.525 | 2.463 | 46.48 | 4.624 |
| **fastPHASE** | 4.100 | 47.59 | 4.595 | 5.316 | 46.97 | 4.439 | 3.426 | 46.56 | 4.113 |
| **MaCH** | - | - | - | - | - | - | 4.096 | 46.39 | 5.967 |

SNPs across individuals. Similarly, phasing is less accurate when the genomic distance increases between SNPs as illustrated in 3.10 as LD reduces with the genomic distance.

**Figure 3.9: Evaluation of linkage disequilibrium error.** The error rate is calculated for SNPs pairs for 5 different ranges of LD (from 0 to 1 with 0.2 step). The plot is generated from 9 datasets (1400 SNPs of 3500 individuals) determined randomly from chromosomes 1, 6, and 17.

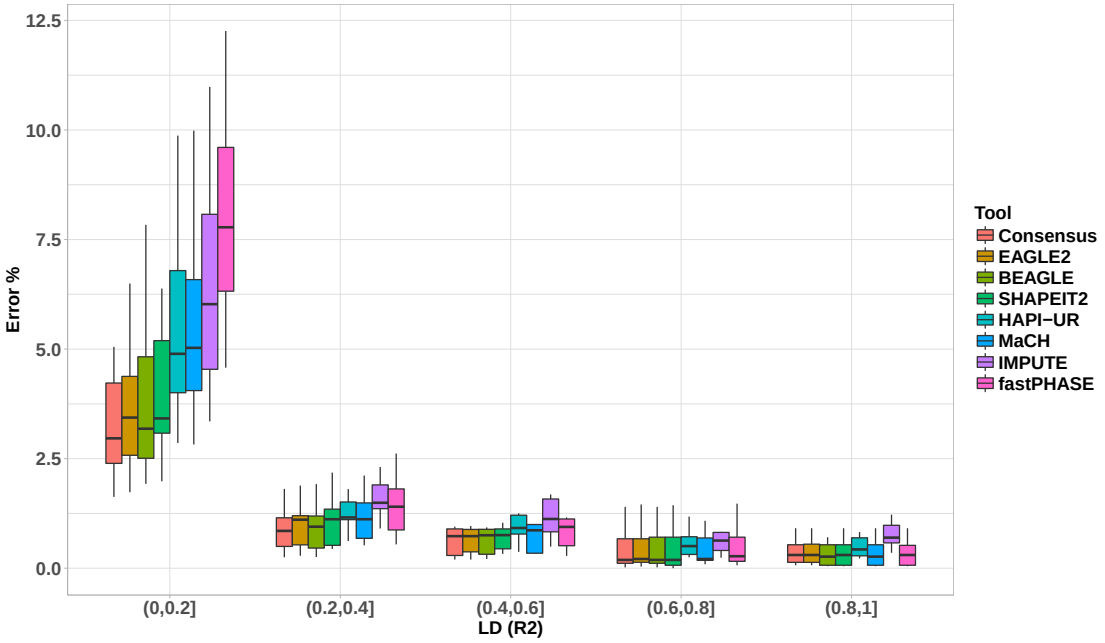
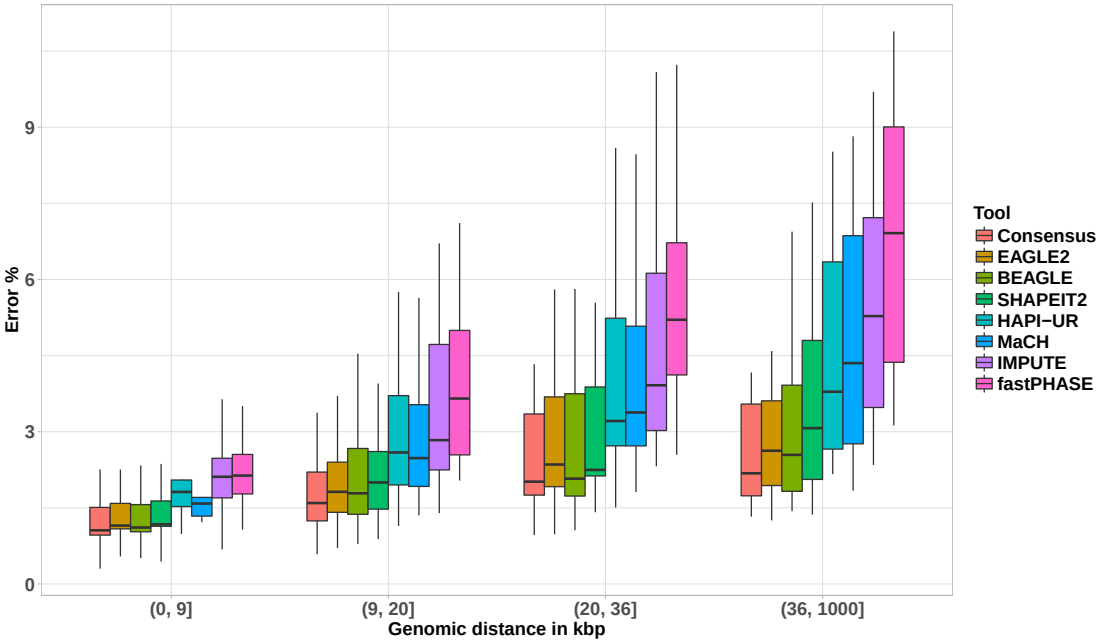


**Figure 3.10: Phasing error rate with respect the genomic distance.** The error rate calculated for SNP pairs within 9 datasets (1400 SNPs of 3500 individuals) determined randomly from chromosomes 1, 6, and 17. This plot is generated for the same pairs used in Figure 3.9. genomic distance ranges are determined using 1st quartile, median, 3rd quartile and maximum distance




### 3.8.3.1 Tool performance

To accurately assess the performance time of the tools, we applied all the tools on the same datasets sequentially using the same machine (Ubuntu 16.04 LTS (Xenial) amd64 - 2 Virtual CPUs AMD Opteron 62xx class 2600 MH - 6 GB RAM). For 6,000 individuals and 1,400 SNPs ( 8 mbp region in chromosome 6): HAPI-UR phased 40,050 SNPs/second, EAGLE2 phased 3,556 SNPs/second, SHAPEIT2 phased 1,785 SNPs/seconds, BEAGLE phased 265 SNPs/seconds, fastPHASE phased 121 SNPs/seconds, and phased MaCH 61 SNPs/seconds. HAPI-UR accomplished the phasing 12 times faster than the closest competitor EAGLE2. The performance time of HAPI-UR was almost linear and slightly affected by the population size, while the performance time of the remaining applied tools almost doubled when doubling the size of the population.

For memory usage, we haven't applied any standard method to assess the space complexity of these methods or to trace the memory usage during the execution, however, BEAGLE was the only tool we struggled with for memory consumption issues, where we had to change the machine to another one with huge memory capacity (64 Gb) to apply it on the whole chromosome of 12,008 individuals.

In this study, all the tools were applied to the whole genomic region with an exception for IMPUTE as the authors suggested to apply it for no longer than 5 Mbp region. Thus for this tool, we divided the long genomic regions into 5 Mbp sub-regions, then we phased each of them separately and concatenated the results in the same way the authors described in the tool's manual.

# Chapter 4

# Evaluation of consensus strategies for haplotype phasing

*This chapter contains the investigations performed to respond to the following research question:*

> *How can we maximise phasing accuracy by aggregating multiple phased haplotypes into a consensus estimator applied to datasets with different characteristics?*

This chapter appears in a manuscript [1] available online through bioRxiv at 10.1101/2020.07.13.175786. It has also been published in Briefings in Bioinformatics - Oxford Academic at this link at doi: 10.1093/bib/bbaa280

[1] Al Bkhetan, Ziad, et al. "Evaluation of consensus strategies for haplotype phasing.". Briefings in Bioinformatics.

## 4.1 Motivation

While a high accuracy of phasing is reported for large dataset with high SNP density (as discussed in the background: 2.4.5 Accuracy of haplotype phasing), changes in dataset characteristics can impact the accuracy dramatically (Browning and Browning, 2011). In the previous chapter, we demonstrated that phasing accuracy can be improved by combining outputs from multiple phasing tools. We also noted that the majority of phasing tools are non-deterministic as they provide different outputs when applied to the same dataset various times. These observations encouraged us to explore how the structure of consensus influences the accuracy of phasing when changing the characteristics of the dataset. Finally, we extended the evaluation of haplotype phasing to one of its major applications by considering the impact of phasing improvement on the downstream genotype imputation. In this chapter, we limited the experiments to the best four phasing tools (SHAPEIT2 (Delaneau et al., 2012), SHAPEIT3 (O'Connell et al., 2016), EAGLE2 (Loh et al., 2016a) and HAPI-UR (Williams et al., 2012)) that are accurate and scalable for large datasets. For genotype imputation, we considered the most popular tools Minimac3 (Das et al., 2018) (used in Michigan imputation servers), pbwt (McCarthy et al., 2016) (used in Sanger imputation servers) as well as BEAGLE5 (Browning et al., 2018).

# Evaluation of consensus strategies for haplotype phasing

Ziad Al Bkhetan [1,4], Gursharan Chana [2], Kotagiri Ramamohanarao [1], Karin Verspoor [1,†], and Benjamin Goudey [3,*,†]

[1]School of Computing and Information Systems, The University of Melbourne, Victoria, Australia.

[2]Department of Medicine, Royal Melbourne Hospital, The University of Melbourne, Victoria, Australia.

[3]IBM Research Australia, Victoria, Australia.

[4]Data61, CSIRO, Canberra, Australia.

## Abstract

**Motivation:** Haplotype phasing is a critical step for many genetic applications but incorrect estimates of phase can negatively impact downstream analyses. One proposed strategy to improve phasing accuracy is to combine multiple independent phasing estimates to overcome the limitations of any individual estimate. As such a strategy is yet to be thoroughly explored, this study provides a comprehensive evaluation of consensus strategies for haplotype phasing, exploring their performance, along with their constituent tools, across a range of real and simulated datasets with different data characteristics and on the downstream task of genotype imputation.

**Results:** Based on the outputs of existing phasing tools, we explore two different strategies to construct haplotype consensus estimators: voting across outputs from multiple phasing tools and multiple outputs of a single non-deterministic tool. We find the consensus approach from multiple tools reduces switch error by an average of 10% compared to any constituent tool when applied to European populations and has the highest accuracy regardless of population ethnicity, sample size, SNP-density or SNP frequency. Furthermore, a consensus provides a small improvement indirectly the downstream task of genotype imputation regardless of which genotype imputation tools were used. Our results provide guidance on how to produce the most accurate phasing estimates and the tradeoffs that a consensus approach may have.

**Availability:** Our implementation of consensus haplotype phasing, consHap, is available freely at `https://github.com/ziadbkh/consHap`.

[†]These authors have contributed jointly to this work as senior authors

# Introduction

Computational haplotype phasing, whereby phase information is statistically estimated from genotype data, remains an essential task in many types of genetic studies such as genome-wide association studies (GWAS) [1], expression quantitative trait loci (eQTL) [2] and genotype imputation [3, 4, 5]. While short-read sequencing technologies can provide phase information [6], studies based on SNP array remain more widely-used, hence there is still a strong demand for statistical approaches for estimating phase. An illustration of this demand is seen in the popular genotype imputation services provided by Sanger [7] and Michigan [4], where the latter has imputed over 58.1 million genomes before July 2020. In the imputations provided by these servers, users select a tool to use for phasing, which will in turn have an impact on the accuracy of all downstream analyses. Given its strong impact as a pre-processing step, it is critical to understand and improve the accuracy of approaches to phasing.

While state-of-the-art phasing tools have an error rate of only a few percent on large, high-density GWAS, error rates are significantly higher in cohorts with low sample size or low-SNP density [8, 9]. This can affect larger studies constructed from multiple smaller cohorts, such as the Haplotype Reference Consortium (HRC) dataset [7], which consists of 38,821 individuals from 20 different cohorts, with mean sample size is 1,630 individuals. Collating such data often requires genotype imputation for study harmonisation which is influenced by phasing accuracy within the small datasets.

One approach that has been proposed to improve phasing accuracy is to combine different estimations of haplotypes into a single consensus estimator through majority voting, a form of ensemble learning [10]. The different haplotypes can be obtained from either several tools [11, 12], which we denote a *multi-tool* consensus, or several iterations of a single but non-deterministic tool [13, 14, 9], which we denote a *multi-iteration* consensus. While these consensus or ensemble approaches have been previously mentioned, there has been little study on which phasing estimates to combine, how to combine them or the impact such techniques have on overall accuracy. Furthermore, preliminary analysis [11] and studies of ensemble methods in the machine learning community [15] highlight that combining multiple models does not necessarily outperform individual tools. Given a consensus estimator has an increased computational cost, requiring a dataset to be phased multiple times, a more rigorous study of this approach is warranted to understand its benefits and trade-offs.

This study provides an evaluation of consensus haplotype approaches. Several combinations of multiple tools or multiple iterations of the same tool are used to construct and evaluate the accuracy of this strategy and the results are compared to the constituent tools. Using data from HapMap III and HRC, we conducted a comprehensive evaluation of the proposed consensus estimators against

state-of-the-art population-based phasing tools and focusing on the ones employed by Michigan and Sanger imputation servers (SHAPEIT2 [16], EAGLE2 [8], SHAPEIT3 [9], and HAPI-UR [14]). We evaluate how the proposed methods are affected by factors known to influence phasing accuracy including sample size, SNP density, minor allele frequency (MAF) and population ethnicity. We explore the impact of pre-phasing on genotype imputation across several imputation tools (pbwt [5] used in Michigan imputation server, Minimac3 [7] used in Sangar imputation server, and Beagle5 [3]). Finally, we analyse the trade-off between runtime and accuracy for the proposed methods and discuss several factors that are likely to mitigate the additional computational cost.

Our results show that consensus haplotype approaches are consistently the most accurate approach for phasing, with computational costs likely to be acceptable in most common scenarios. We also find that consensus haplotype phasing leads to the most accurate imputation compared to most combinations of phasing and genotype imputation tools available on Sanger and Michigan imputation servers, albeit with only modest increases over individual tools. SHAPEIT2 and EAGLE2 provide the most accurate phasing from any individual phasing tool, with EAGLE2 provided better genotype imputation in our evaluation. Our study results provide guidance on the strengths and limitations of consensus strategies for haplotype phasing that can help other researchers to conduct accurate genetic investigations and highlight several directions for future research.

# Methods

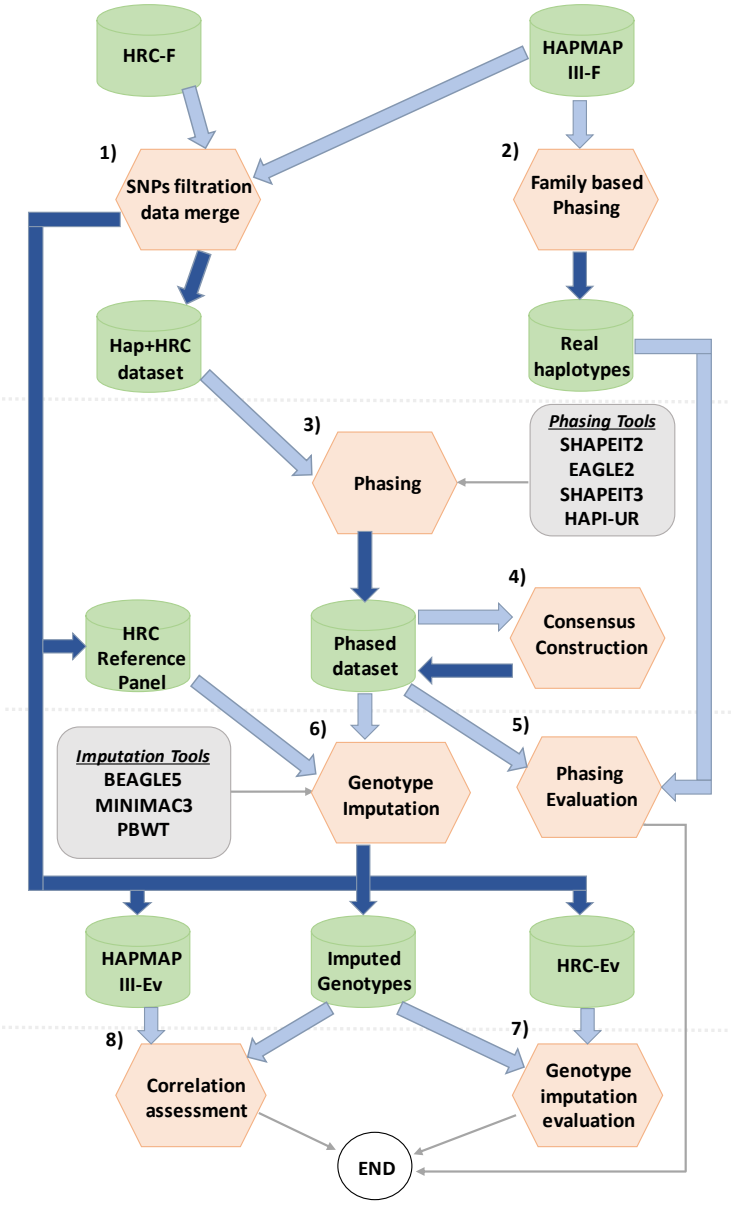
## Study workflow

Figure 1 illustrates the applied analysis workflow: **1)** Datasets preparation. **2)** Real haplotypes preparation. **3)** Phasing is applied to the Hap+HRC dataset using all individual phasing tools, then results are combined into a consensus estimators in 4). **5)** Estimated HAPMAP III Trios are evaluated against real haplotypes using switch error. **6)** The SNPs of the reference panel are imputed to the phased Hap+HRC dataset using three genotype imputation tools. **7)** Imputed genotypes are evaluated against the experimental genotypes filtered in step (1). **8)** Correlation assessment for phasing and genotype imputation accuracy at each individual scale (same individuals in step (5)).

## Datasets and preparation

We analyzed data from HapMap phase III [17] and the Haplotype Reference Consortium (HRC) [7] with quality control for these datasets described in their respective publications. The majority of the analysis focuses on a combination of European samples from HapMap III and HRC datasets, which we denote the *Hap+HRC* dataset. We make use of 8,259 individuals of European descent, combining 69 Utah residents with Northern and Western European ancestry (CEU) obtained from HapMap phase III and 8,189 British samples from HRC (specifically the UK10UK and IBD cohorts).

Figure 1: Workflow applied in this study. Dark blue arrows represent saving data into files, while light blue arrows represent reading the saved data. Thin arrow used to show additional information or the end of the workflow. Cylinders are used to represent data storage. Hexagons are used to represent applied procedures.



HRC and HapMap samples were combined, keeping SNPs with MAF $> 0.05$ and are common to the two datasets. As both studies have very high density, we use LD-based pruning via PLINK [18] (*–indep-pairwise* with $r^2 = 0.95$), resulting in a SNP density similar to the UK Biobank dataset as per [8]. Hap+HRC individuals with the original SNP density were saved in *HAPMAP III-Ev* and *HRC-Ev* for genotype imputation evaluation. All analysis for the Hap+HRC dataset focus on 5 chromosomes, totalling 151,278 SNPs (50,521 on chr2, 38,930 on chr6, 30,580 on chr16, 21,928 on chr6 and 9,319 on chr21). Phasing accuracy of the Hap+HRC dataset is benchmarked using 52 trios in the CEU dataset (denoted *Real haplotypes*), where phase is resolved using parental genotypes. As the HRC dataset consists of statistically estimated haplotypes, it could not be used for evaluation. HRC samples excluded from the Hap+HRC dataset are used as a reference panel for genotype imputation.

We additionally use all 838 samples (excluding any parental samples) from the HapMap phase III to emulate phasing a multi-ethnic cohort, using the same SNPs of chromosome 21 of the Hap+HRC dataset. Evaluation is conducted using trios from five populations: African ancestry in Southwest USA (ASW), Utah residents with European ancestry (CEU), Mexican ancestry in Los Angeles (MEX), Maasai ancestry in Kenya (MKK), and Yoruba in Ibadan, Nigeria (YRI). The number of trios in each population is reported in Table 2.

## Haplotype phasing and genotype imputation tools

Haplotype phasing was carried out using four tools: SHAPEIT2 (v2.r904), EAGLE2 (v2.4.1), HAPI-UR (v1.01), and SHAPEIT3 (v3.r881), with the first three tools motivated by their availability as part of the widely-used Sanger and Michigan genotype imputation servers. Phasing tools were applied using their default parameters unless otherwise was noted. SHAPEIT3 was applied with fast flag enabled as mentioned in its documentation. When phasing small datasets via SHAPEIT3, the cluster size parameter was set to (2× sample size - 1), a mandatory setting when the sample size is below 4,000.

We utilised the following genotype imputation tools applied with default parameters: Minmac3 (2.0.1), used in the Michigan imputation servers, pbwt (3.0), used in the Sanger imputation servers, and Beagle v5.

## Consensus estimator

Rather than making use a single model, a consensus approach is a form of ensemble learning, whereby the predictions from multiple independent models are combined to obtain better predictive performance than any of the constituent models [10]. In the context of haplotype phasing, existing phasing algorithms are usually based on Hidden Markov Models, with the different tools each making different assumptions and implementing different heuristics to improve runtime [19, 13].

By combining the predictions of these models, it is hoped that a majoity vote can overcome the limitations and biases of each individual model, resulting in a higher accuracy on average.

IN this work, we focus on consensus approach, demonted *consHap*, based on a majority voting approach described in Algorithm 1 and previously proposed in [11, 20]. The method takes $M$ phasing estimates where the output from the $j$-th tool, denoted $h_j$, is a binary array of length $P$, representing the sequence of alleles at every heterozygous SNP on a single homologous chromosome. $h_j[i]$ denotes whether the $i$-th heterozygous SNP carries a major or minor allele (0 or 1 respectively). Only one homologous chromosome needs to be represented, with the remaining copy being the logical complement. For each SNP, our method takes a majority vote across all input phasing estimates. Any method that is not concordant with the majority vote has phasing estimates for remaining SNPs flipped (i.e we take the logical complement) in order to align the phase of the remaining estimates with those of the majority.

**Algorithm 1:** Consensus haplotype estimator algorithm.

**Input:** $\{h_1..h_m\}$ – $M$ phasing estimates, binary vectors of size $P$
**Output:** *cons* – binary vector of size $P$
**for** $i \leftarrow 0$ *to* $P$ **do**
  $cons[i] \leftarrow majority\_vote(h_1[i], \cdots, h_j[i])$;
  **for** $j \leftarrow 0$ *to* $n$ **do**
    **if** $i+1 \leq P$ *and* $h_j[i+1] \neq cons_j[i+1]$ **then**
      $h_j[i+1:P] \leftarrow complement(h_j[i+1:P])$
    **end**
  **end**
**end**

We consider two different forms of consensus estimators, either multi-tool approaches that combine the output of different phasing tools, or multi-iteration approaches that combine the output of multiple iterations of one non-deterministic tool. We also explore the impact of combining these approaches i.e. multiple iterations from multiple tools. Most phasing tools considered in this work are non-deterministic, in the sense that they produce different outputs when running with the same data. Only EAGLE2 produces deterministic results but can be forced to emulate non-deterministic output by permuting the order of individuals in the input [11].

We use a consistent naming convention to describe a particular implementation of consHap throughout this work. Individual tools are abbreviated ($S_2$: SHAPEIT2, $S_3$: SHAPEIT3, $E_2$: EAGLE2, $H_r$: HAPI-UR) and concatenated together if multiple tools are used. If multiple outputs from a single tool are used, we denote the number of iterations in parentheses. Thus consHap-$S_2E_2S_3$ is based on the output of SHAPEIT2, EAGLE2, and SHAPEIT3 while consHap-$S_2$*(3)* is based on three different iterations of SHAPEIT2.

### Evaluation criteria

The evaluation of phasing accuracy is based primarily on the proportion of switch errors (SE), which occur when the phase allocated to a heterozygous variant is incorrectly switched relative to the prior heterozygous [13]. While switch error rate is primarily reported as an average across all samples in a dataset, we also consider the variation of switch error at an individual level. This also enables us to evaluate there is a statistically significant difference between a pair of phasing tools via a one-sided binomial test, with the number of trials as the number of individuals and the probability of an improvement as 0.5 given the null hypothesis that two tools have the same accuracy on average.

Evaluation of phasing performance with varying SNP density was conducted on five subsets of the Hap+HRC datasets, restricted to chromosomes 16 and 21, and thinned to a SNP density ranging from 200SNPs/1Mb to 1000 SNPs/Mb. This pruning (via PLINK's *–bp-space* command) removed a random SNP from any possible pair if the distance was less than the desired resolution. Similarly, five datasets with different population sizes were generated by a random selection of samples from the Hap+HRC dataset, again restricted to chromosomes 16 and 21. The error rate of phasing as MAF varies was calculated across all chromosomes of the Hap+HRC dataset.

Using the Hap+HRC dataset that was thinned for the phasing evaluation, we imputed all pruned SNPs using the reference panel into the Hap+HRC dataset, a masking experiment similar to [21]. Concordance of imputed and actual genotype calls was measured using Pearson correlation between the allele dosage and the sum of the posterior allele probabilities ($r_a^2$) for each SNP [3, 4].

## Results

### Consensus performance across European cohort

We first evaluated consensus haplotype approaches, which we refer to as *consHap*, across five chromosomes of the Hap+HRC dataset and compared the proportion of switch errors (SE) with that achieved by several state-of-the-art phasing tools. We focus on two variants: a multi-tool consensus of SHAPEIT2, EAGLE2 and SHAPEIT3 (consHap-$S_2E_2S_3$) and a multi-iteration consensus of three outputs of SHAPEIT2 (consHap-$S_2(3)$).

Table 1 shows that the consensus estimators, consHap-$S_2E_2S_3$ and consHap-$S_2(3)$, improved on the most accurate individual tool across the evaluated chromosomes of the Hap+HRC dataset by an average of 9.8% (consHap-$S_2E_2S_3$) and 5.6%(consHap-$S_2(3)$). The multi-tool consHap-$S_2E_2S_3$ consistently improves upon the multi-iteration consHap-$S_2(3)$ by between 2.3 and 8.8% depending on the chromosome. SHAPEIT2 is the most accurate individual tool in all cases, showing consistent improvements beyond EAGLE2. The average length of correctly phased haplotype blocks obtained by consHap-S2E2S3 and consHap-S2(3) is 421, and 403 SNPs respectively, representing 45 and 27

Table 1: Switch error % and total number of switches for evaluated phasing tools on 52 individuals for five chromosomes of the Hap+HRC dataset. Bolded values represent the lowest error while underlined values have the lowest switch error for any individual tool (i.e excluding consensus). Improvement (bottom row) highlights the percentage of switch error improvement between consHap-$S_2E_2S_3$ and the best individual tool.

| Tool | Chr2 | Chr6 | Chr11 | Chr16 | Chr21 | Switches |
|---|---|---|---|---|---|---|
| **consHap-$S_2E_2S_3$** | **0.86** | **0.83** | **0.92** | **1.10** | **1.23** | **18,763** |
| **consHap-$S_2(3)$** | 0.88 | 0.85 | 0.95 | 1.14 | 1.32 | 19,392 |
| **SHAPEIT2** | <u>0.95</u> | <u>0.92</u> | <u>1.00</u> | <u>1.20</u> | <u>1.40</u> | <u>20,742</u> |
| **EAGLE2** | 0.99 | 0.94 | 1.04 | 1.24 | 1.41 | 21,352 |
| **SHAPEIT3** | 1.1 | 1.06 | 1.16 | 1.38 | 1.51 | 23,850 |
| **HAPI-UR** | 1.51 | 1.42 | 1.61 | 1.89 | 2.12 | 32,604 |
| **Improvement** | 9.76% | 10.38% | 8.15% | 8.52% | 11.72% | 1,979 |

SNPs longer blocks compared to the best individual tool as shown in supplementary Table 3.

Rather than looking at average switch error, we can also consider how often a given phasing tool results are the most accurate phasing of an individual. Supplementary Table 2 shows that consHap-$S_2E_2S_3$ and consHap-$S_2(3)$ were more accurate than all individual tools for 76% and 56% of individuals, representing a significant improvement in performance (one-sided binomial test $p < 5 \times 10^{-5}$ for both consensus methods). The scenario where individual tools perform better than the consensus arises when the majority of phasing outputs in the consensus are in error but out-vote the more accurate minority.

We also explored a range of other consensus configurations (Supplementary Table 4 and 5). The highest accuracy was obtained by consHap-$S_2(8)E_2(7)$, a consensus of 8 iterations of SHAPEIT2 combined with 7 iterations of EAGLE2, showing a mean SE of 1.057% compared to 1.168% by consHap-$S_2E_2S_3$, but with a cost of a significant increase in runtime. Supplementary Table 4 shows that consensus estimators constructed from any three tools obtained higher accuracy than any of the tools individually. However, Supplementary Figure 1 shows this was not the case with a multi-iteration consensus, where multiple iterations of SHAPEIT3 or HAPI-UR failed to outperform a single iteration of SHAPEIT2 or EAGLE2.

## Performance across factors that influence accuracy

We evaluated consHap-$S_2E_2S_3$ and consHap-$S_2(3)$ over several subsets of the Hap+HRC dataset to assess their robustness against factors known to influence phasing accuracy including SNP density, population size, and minor allele frequency (MAF) [22, 13].

As shown in Figure 2, consHap, in both multi-tool and multi-iteration configurations, outperforms all individual tools regardless of SNP density from subsampled versions of the Hap+HRC

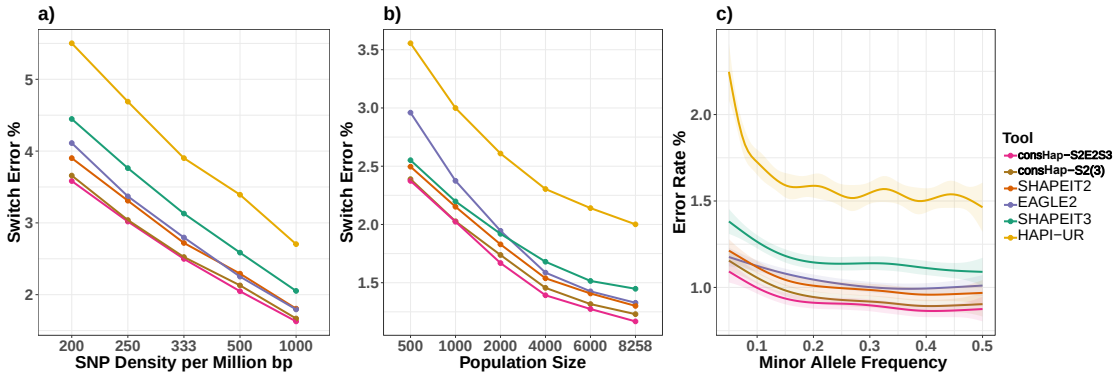



Figure 2: Robustness of the consensus approach to a) SNP density, b) population size and c) minor allele frequency (MAF), as measured by switch error and Error rate. c) is generated from the Hap+HRC dataset and smoothed using a generalized additive model (GAM), while a) and b) are generated by manipulating the data of chromosomes 16 and 21 from the Hap+HRC dataset to obtain the desired characteristics as explained in supplementary materials.

dataset. Overall, the average SE increases by an average of 1.94% to 4.2% across all approaches as the SNP density is reduced. When changing the resolution of SNPs from 200SNPs/Mb to 1000SNPs/Mb, as shown in Figure 2 (a), we see consHap-$S_2E_2S_3$ and consHap-$S_2(3)$ improved the accuracy by an average of 8.54% and 6.65% respectively compared to the best single tool. Interestingly, the best individual tool changes from SHAPEIT2 to EAGLE2 as the density of SNPs gets higher than 500SNPs/Mb. We again observe that the multi-tool consensus, consHap-$S_2E_2S_3$, outperforms the multi-iteration consensus, consHap-$S_2(3)$ showing a mean improvement of 2%.

Similar results are seen when varying sample size. In Figure 2 (b) we see consHap-$S_2E_2S_3$ and consHap-$S_2(3)$ improve accuracy by 7.7% and 4.4% when changing the population size from 500 to 8,258 individuals, the percentage improvement increases as sample size increases from 4.8% to 11.7% and 3.5% to 5.4% for consHap-$S_2E_2S_3$ and consHap-$S_2(3)$, respectively.

Furthermore, consHap obtained the minimal error rate regardless of the minor allele frequency (MAF) as illustrated in Figure 2 (c). There was 9.7% (averaged for all MAF ranges) error reduction obtained by consHap-$S_2E_2S_3$ compared to the best individual tool. While the accuracy of all phasing tools decreases as we examine increasingly rare SNPs, we observe that while SHAPEIT2 is the most accurate individual tool for variants with MAF $> 0.1$, its accuracy drops below that of EAGLE2 once when MAF $< 0.1$.

## Accuracy gains of increasing consensus iterations

The results in previous sections consider consHap-$S_2(3)$, a specific instance of consHap-$S_2(n)$, a consensus estimator based on $n$ iterations of SHAPEIT2. We can evaluate the impact on perfor-

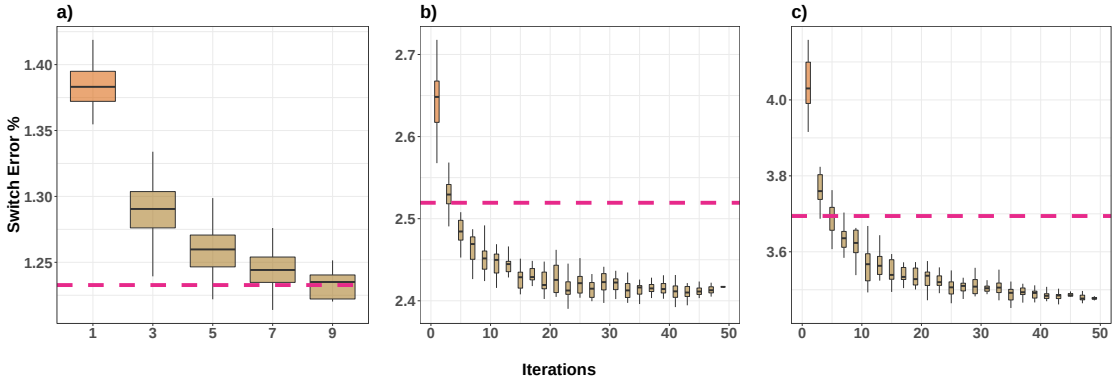



Figure 3: Performance of multi-iteration consensus for varying numbers of iterations over chromosome 21 of the Hap+HRC dataset a) without modification, b) reduced to 500 individuals and c) with SNP density reduced to 1 SNP per 5kb. The number of iterations is shown on the x-axis. The dashed pink line shows switch error for consHap-$S_2E_2S_3$.

mance as the parameter for iterations, $n$, is increased. Moreover, as the consensus relies on the non-deterministic output of SHAPEIT2, we can also examine the distribution of accuracy from different runs of this tool. This is achieved by running SHAPEIT2 up to 50 times and, for each $n$ of interest, selecting $n$ random iterations as inputs into the multi-iteration consensus. This selection is repeated 10 times, to derive the distribution of error.

Initially, we examine the performance of consHap-$S_2(n)$ as $n$ is increased, focusing on only chromosome 21 of the Hap+HRC dataset. Figure 3 (a) highlights a 7.2% improvement in accuracy when shifting from 1 iteration to 3 iterations (mean SE from 1.39% to 1.29%, respectively) with 9 iterations required to achieve the same mean SE as the multi-tool consHap-$S_2E_2S_3$ (SE = 1.23%). Additional iterations continue to reduce SE with diminishing returns, with details of the accumulative improvement is reported in Supplementary Table 6. Similar observations are noted when constructing a multi-iteration consensus from other tools (Supplementary Figure 1).

As datasets with low sample size or SNP density can be phased faster, we hypothesised that a multi-iteration consensus combining a large number of phasing outputs may outperform a multi-tool consensus. Evaluating up to 49 iterations on a data set with reduced sample size (n=500), we found that consHap-$S_2(3)$ obtained similar results to consHap-$S_2E_2S_3$ (SE = 2.53%) as illustrated in Figure 3 (b). Similar results are observed when SNP density is reduced but sample size is maintained, as shown in Figure 3 (c). Here, consensus from five iterations of SHAPEIT2 or more improve beyond consHap-$S_2E_2S_3$ (SE = 3.7%). While these results were produced for chromosome 21, Supplementary Figure 2 shows similar findings on chromosome 16. Supplementary Figure 3 shows these results hold as population size and SNP density are varied further.

To understand why these multi-iteration consensuses improve beyond consHap-$S_2E_2S_3$, we in-

vestigated the individual tools used in the consensus construction. Consistent with results in Figure 2, Supplementary Figure 4 shows EAGLE2 has a relatively high SE over datasets with small sample size (SE= 3.2% compared to 2.64% and 2.69% for SHAPEIT2 and SHAPEIT3, respectively) that may impact the performance of consHap-$S_2E_2S_3$ in Figure 3 (b). When SNP density is reduced, SHAPEIT3 obtained SE = 4.6% compared to 4.17% and 4.04% for EAGLE2 and SHAPEIT2 respectively, which may impact consHap-$S_2E_2S_3$ in Figure 3 (c). These analyses indicates that SHAPEIT2 has the most robust performance across low sample and low SNP density datasets. Additionally, we found that constructing a consensus using several iterations of one non-deterministic tool is not guaranteed to outperform all individual tools, with Supplementary Figure 5 showing multiple iterations of HAPI-UR failing to outperform a single iteration of SHAPEIT2 or EAGLE2, given the large gap in accuracy.

## Phasing performance on a multi-ethnic cohort

We evaluate the performance of phasing on a multi-ethnic cohort formed by combining all samples from the HapMap. As shown in Table 2, we find consHap-$S_2E_2S_3$ outperforms all individual tools, with improvements ranging from 4.2% (CEU) to 11.2% (MKK). consHap-$S_2E_2S_3$ also outperforms consHap-$S_2(3)$ across all populations, though the improvement is only 1.15% on average.

As the smaller sample size used for this experiment means error rates across all phasing tools are typically high (3-4%) for all ethnicities, we see a lower error rate in the Maasai in Kinyawa, Kenya (MKK) cohort, likely because this is the largest population in this multi-ethnic dataset. We also observe a substantially higher error rate for the CEU trios compared to Table 1, due to both the smaller number of samples used for phasing and the non-European nature of this experiment.

Given the results from Figure 3 showing that higher iterations of a single tool had stronger results than a multi-tool consensus, we repeated the analysis with 15 iterations of SHAPEIT2 (consHap-$S_2(15)$). The resulting error rates were lower than all other phasing estimators, with the exception of the YRI population where consHap-$S_2E_2S_3$ was the most accurate approach.

## Impact of different phasing strategies on genotype imputation

Evaluations of phasing reported in this work consistently show the improvements obtained by consHap. However, we can also assess the impact of consHap on downstream tasks. We consider the use of consHap as well as all phasing tools for pre-phasing for genotype imputation. As SNPs were filtered out of Hap+HRC , we imputed them again and compared them to the real SNPs obtained experimentally with the original dataset. Genotype imputation tools (Beagle5, Minimac3, and pbwt) are applied to the Hap+HRC dataset after phasing by all methods described so far, as well as considering the data unphased.

The results in Table 3 for SNPs within three ranges of MAF show imputation using Beagle5

Table 2: Switch error % when phasing different populations from the HapMap III project. The title of each column shows abbreviated population names and number of trios (in parentheses), with the full populations names in Section . Populations information are in supplementary methods. The approaches are sorted based on the average switch errors. Bolded results are the lowest switch error. Italic results are the best across all tests. Underlined results are the best individual tool. The improvement row demonstrates the percentage improvement between consHap-$S_2E_2S_3$ and the best single tool.

| Tool | ASW (30) | CEU (52) | MEX (27) | MKK (28) | YRI (54) | Mean |
|---|---|---|---|---|---|---|
| **consHap-$S_2E_2S_3$** | 4.480 | 3.528 | 3.929 | 1.073 | **4.096** | 3.421 |
| **consHap-$S_2(3)$** | 4.552 | 3.560 | 3.951 | 1.032 | 4.210 | 3.461 |
| **SHAPEIT2** | 4.77 | 3.684 | 4.149 | 1.208 | 4.430 | 3.648 |
| **SHAPEIT3** | 4.816 | 3.744 | 4.237 | 1.266 | 4.418 | 3.696 |
| **EAGLE2** | 6.343 | 4.330 | 4.527 | 2.044 | 5.556 | 4.560 |
| **HAPI-UR** | 7.778 | 5.011 | 5.446 | 4.966 | 7.844 | 6.209 |
| **consHap-$S_2(15)$** | **4.374** | **3.448** | **3.883** | **0.994** | 4.110 | **3.362** |
| **Improvement** | 6.1% | 4.24% | 5.28% | 11.2% | 7.29% | 6.82% |

consistently obtained the highest accuracy, followed by Minimac3 and pbwt. With respect to phasing tools, consHap, regardless of configuration, obtained higher accuracy than any individual tool, with the best results obtained by consHap-$S_2(8)E_2(7)$. Compared to the results obtained by the best individual tool for pre-phasing, consHap-$S_2(8)E_2(7)$ and consHap-$S_2E_2S_3$, correctly imputed a further 2,800,715 and 1,266,483 genotypes respectively, which equates to 350 correctly imputed SNPs for each individual, within 5 chromosomes. The highest relative improvement was for SNPs with MAF in the range from 0.5% to 5%. With regard to pre-phasing performed via a single tool, EAGLE2 had the highest accuracy in all but two cases where SHAPEIT2 was better. Detailed results showing further imputation metrics are shown in Supplementary Tables 7-9.

Given the results in Table 3 appear to indicate that improved phasing accuracy may have little improvement in genotype imputation, we sought to quantify this relationship. The correlation between phasing and genotype imputation accuracy was evaluated for the 52 individuals of HapMap III, as these individuals have both known real haplotypes and high-density genotype data. Results in Supplementary Figures 6 and 7 show a positive correlation between phasing and imputation accuracy with an mean Pearson correlation across all individuals of $r$=0.35 (range: 0.26 to 0.48), a highly significant relationship (mean $p = 6 \times 10^{-6}$, range: $6 \times 10^{-16}$ to $3 \times 10^{-5}$). Overall, Beagle5 was more influenced by phasing accuracy compared to other imputation tools mean $r = 0.4$, mean $p = 4 \times 10^{-9}$), then Minimac3 (mean $r = 0.35$, mean $p = 9 \times 10^{-7}$) and finally pbwt (mean $r = 0.3$, mean $p = 7.1 \times 10^{-6}$). Detailed results are reported in Supplementary Table 10.

Table 3: Genotype imputation evaluation. $r^2$ summarised for 54,230,334,071 SNPs across three MAF ranges: 0.01% to 0.5%, 0.5% to 5%, and 5% to 50%. Improvement column represents the accumulative additional correctly imputed SNPs compared to the pre-phasing done by the phasing tool in the next row. For example, the additional correctly imputed SNPs by consHap-$S_2(8)E_2(7)$ in the first row is 1,447,183 SNPs compared to consHap-$S_2E_2S_3$ and (1,447,183 + 560,920) SNPs compared to consHap-$S_2(3)$. Bold results indicate the highest accuracy for each imputation tool, underlined results indicate the most accurate individual phasing tools.

| | Phased by | [0.01%,0.5% ] | (0.5%, 5%] | ¿ 5% | Improvement |
|---|---|---|---|---|---|
| **Imputed by Beagle5** | **consHap-$S_2(8)E_2(7)$** | **0.314** | **0.768** | **0.972** | 1,447,183 |
| | **consHap-$S_2E_2S_3$** | 0.309 | 0.766 | 0.971 | 560,920 |
| | **consHap-$S_2(3)$** | 0.308 | 0.765 | 0.971 | 723,566 |
| | **EAGLE2** | <u>0.307</u> | <u>0.763</u> | <u>0.971</u> | 254,420 |
| | **SHAPEIT2** | 0.306 | <u>0.763</u> | <u>0.971</u> | 2,817,004 |
| | **SHAPEIT3** | 0.299 | 0.758 | 0.97 | 3,740,168 |
| | **HAPI-UR** | 0.296 | 0.751 | 0.969 | 1,032,159,933 |
| | **unphased** | 0.066 | 0.327 | 0.745 | - |
| **Imputed by Minimac3** | **consHap-$S_2(8)E_2(7)$** | **0.315** | **0.764** | **0.971** | 1,473,243 |
| | **consHap-$S_2E_2S_3$** | 0.31 | 0.762 | 0.97 | 545,408 |
| | **consHap-$S_2(3)$** | 0.309 | 0.761 | 0.97 | 660,184 |
| | **EAGLE2** | <u>0.308</u> | <u>0.759</u> | <u>0.97</u> | 285,388 |
| | **SHAPEIT2** | 0.307 | <u>0.759</u> | <u>0.97</u> | 2,887,572 |
| | **SHAPEIT3** | 0.299 | 0.754 | 0.969 | 355,6971 |
| | **HAPI-UR** | 0.297 | 0.746 | 0.969 | 980,415,942 |
| | **unphased** | 0.065 | 0.298 | 0.759 | - |
| **Imputed by pbwt** | **consHap-$S_2(8)E_2(7)$** | **0.279** | **0.739** | **0.969** | 1,682,270 |
| | **consHap-$S_2E_2S_3$** | 0.274 | 0.736 | 0.968 | 723,411 |
| | **consHap-$S_2(3)$** | 0.272 | 0.735 | 0.968 | 585,959 |
| | **EAGLE2** | <u>0.271</u> | <u>0.734</u> | <u>0.968</u> | 541,770 |
| | **SHAPEIT2** | 0.27 | 0.733 | <u>0.968</u> | 3,281,317 |
| | **SHAPEIT3** | 0.261 | 0.727 | 0.967 | 4,157,745 |
| | **HAPI-UR** | 0.259 | 0.719 | 0.966 | 1,140,380,286 |
| | **unphased** | 0.047 | 0.249 | 0.725 | - |

## Runtime and computational cost

The improved accuracy of the consensus comes at a cost of increased runtime. Figure 4 shows the runtime of the individual tools used in this analysis when evaluated across chromosome 21 of the Hap+HRC dataset, as the SNP density and sample size are varied. When phasing tools are applied in parallel, the total time required by consHap can be calculated as $max(t_1, t_2, ..., t_n)+t_0$ where $t_n$ is phasing time of the tool $n$ and $t_0$ is the time required to aggregate the results together. If the tools

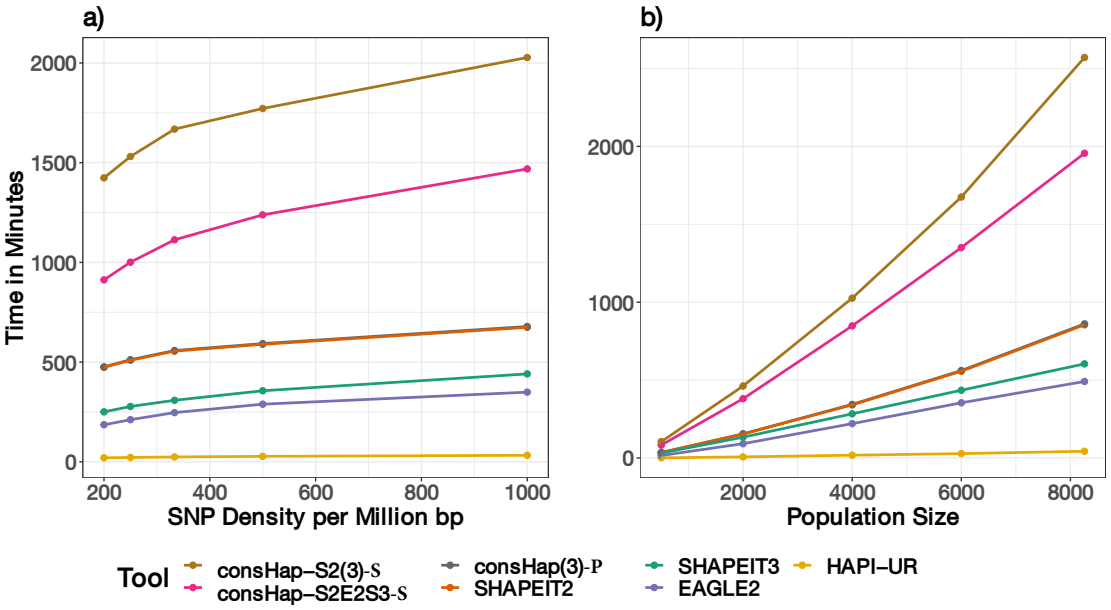



Figure 4: The execution time of phasing tools with respect to population size in (a) and SNP density in (b). Note we include timings for consHap when tools are run in serial (-S suffix) or parallel (-P suffix). Subsampling of SNPs and samples was performed as per Figure 3.

are applied sequentially, the total time can be calculated as $t_{total} = \sum_{i=1}^{n} t_i + t_0$. For chromosome 21 of the Hap+HRC dataset, it required 32.2 and 42.6 hours to construct consHap-$S_2E_2S_3$ and consHap-$S_2(3)$ considering the worst-case scenario where phasing is applied sequentially and using one thread. This runtime reduces dramatically when phasing tools can be run in parallel to 14.2 hours for both consHap-$S_2E_2S_3$ and consHap-$S_2(3)$, with SHAPEIT2 as the bottleneck. The time to conduct voting across the tools was 1.6% of the time to run the tools in parallel (6 minutes). Individually, HAPI-UR was the fastest phasing tool, with a required execution time around 11 times faster than the closest competitor EAGLE2, while SHAPEIT2 is the slowest. Most tools scale linearly with the number of samples, with the exception of SHAPEIT2, while the high speed of phasing for HAPIUR compared to other approaches is demonstrated across all SNP densities and sample sizes.

## Discussion

In this study, we evaluate the performance of consensus haplotype estimators that use majority voting to combine multiple phasing outputs from different phasing tools or multiple iterations of a

single non-deterministic tool. This study provides the first comprehensive evaluation of consensus phasing approaches, demonstrating that the proposed consensus approach leads to stronger phasing accuracy compared to using a single tool and it is robust across many characteristics known to impact phasing performance. We also assess the impact of phasing accuracy on genotype imputation, one of the major application of phasing, across all phasing and genotype imputation tools available in the widely-used Sanger and Michigan imputation servers, with the resulting findings providing guidance for researchers using these services.

In our evaluations, a consensus from multiple phasing tools almost always has higher accuracy than a consensus from the multiple outputs from the same tool. The consensus estimators are a form of ensemble learning whereby multiple predictive models are combined to produce a single prediction. Ensemble models typically have the strongest improvement above their constituent predictors when each model provide accurate but independent predictions [10]. With phasing, we would intuitively imagine that phasing outputs from the same non-deterministic tool will be less independent than outputs from different tools. Only across datasets with low sample size or low SNP density did we observe improvements from combining multiple outputs of SHAPEIT2 compared to multi-tool consensus. Even here, we required at least 5 independent runs of SHAPEIT2 to see such improvements and improvements in SE remained less than 0.2%.

Our study also builds upon previous benchmarks of phasing tools to characterize the performance of the constituent tools. As with previous studies, we find no individual tool is the best in all scenarios. SHAPEIT2 and EAGLE2 provided the strongest results, while HAPI-UR performs consistently worse, as previously reported [11, 20]. Also consistent with prior results [8], SHAPEIT3 tended to provide somewhat less accurate results than SHAPEIT2 or EAGLE2 but had a reduced runtime and still had higher accuracy than HAPI-UR. One interesting observation was that EAGLE2 displayed a substantially sharper drop in accuracy when sample size was low (less than 2000 individuals). This may be due to EAGLE2 use of long-range phasing to generate an initial haplotype set when no reference panel is provided, but if only relatively few samples are present, it may be hard to find individuals who share haplotypes [13].

The improvements in phasing accuracy achieved through a consensus approach do so at the cost of extra computational overhead. However, there are three complementary aspects related to phasing runtime that mitigate the computational overhead incurred by our consensus approach. First, the strong focus on improving run-time for many phasing algorithms has led to computational complexity reducing from quadratic in the number of samples to linear [19]. Secondly, access to high-throughput computational resources, whether via local high-throughput compute servers or whether via cloud services, are being more accessible, enabling the parallel phasing of data. If the running of multiple phasing tools can be parallelized, the observed runtime will be approximately the runtime of the slowest individual tool. Finally, phasing typically only needs to be conducted once for a study as opposed to being an analysis that needs to be run multiple times, and hence

the once-off additional costs, may not be substantial compared to its re-use as part of downstream analyses. While performance time reported in the results of this work represent the worst-case scenario where phasing was applied sequentially using one thread on an average PC, these three facets of computational performance indicate consensus phasing is likely to be a practical option for most datasets.

Our study makes an important contribution in demonstrating the impact of phasing error on genotype imputation. In line with previously reported results [4, 3], we found that genotype imputation tools obtain similar accuracy, with Beagle5 consistently reporting the highest accuracy regardless of variant allele frequency. For each imputation tool, we found that the improvement of phasing accuracy reported for the consensus estimators led to the most accurate genotype imputation with 2,800,715 and 1,266,483 additional correctly imputed SNPs for 8,189 individuals within 5 chromosomes when using consHap-$S_2(8)E_2(7)$ and consHap-$S_2E_2S_3$, respectively compared to the best individual tool. Errors in imputed genotypes of SNPs have been reported to increase false positives and negatives in downstream analysis [23]. Therefore, the improvements in accuracy obtained by the consensus estimator will contribute to more accurate genetic analyses.

A limitation of our phasing evaluation is that it was carried out on at most 191 individuals from HapMap III, with most analyses focusing on the subset of 52 CEU individuals. However, known haplotype data is not readily available in many large studies and many previous empirical evaluations also suffer from this constraint [13, 8, 9, 11]. We restricted our experiments to a set of phasing and imputation tools (using their default parameters) available in Michigan and Sanger imputation servers given their widespread usage and demonstrated utility. While this limitation means we do not explore some of the most recently published phasing tools, including SHAPEIT4 [6] and BEAGLE5 [3], the tools have tended to improve scalability substantially but have modest impacts on accuracy. As the computational complexity of phasing algorithms drops, we believe the case for consensus approaches will be strengthened, given increased runtime is the strongest drawback.

While the results in this work show the promise of consensus approaches to phasing, future investigations on how best construct a consensus estimator, such as weighting the output of each tool depending on the properties of the data being phased, may yield even stronger improvements in performance. It may also be useful to explore consensus constructed from more diverse tools, with novel approaches such as the random-forest based phasing provided by [24], potentially resulting in greater independence between inputs and hence likely producing a more accurate phasing result. And finally, it may also be possible to alter the optimization of existing HMM-based phasing tools so that they produce multiple local optima, rather than the single best solution, and take a vote over these without requiring multiple independent runs of the tool and hence producing a census result without the runtime cost.

## Acknowledgements

The Haplotype Reference Consortium (HRC) dataset is used in a form agreed by The University of Melbourne with Wellcome Trust Sanger Institute.

## Funding

This work was supported by MRS scholarship [103500], the University of Melbourne and a top-up scholarship, Data61 awarded to ZB.

## References

[1] Brian L Browning and Sharon R Browning. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(5):365–375, 2007.

[2] Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J Topol, and Nicholas J Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215, 2011.

[3] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[4] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284, 2016.

[5] Richard Durbin. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272, 2014.

[6] Olivier Delaneau, Jean-François Zagury, Matthew R Robinson, Jonathan L Marchini, and Emmanouil T Dermitzakis. Accurate, scalable and integrative haplotype estimation. *Nature communications*, 10(1):1–10, 2019.

[7] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279, 2016.

[8] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443, 2016.

[9] Jared O'Connell, Kevin Sharp, Nick Shrine, Louise Wain, Ian Hall, Martin Tobin, Jean-Francois Zagury, Olivier Delaneau, and Jonathan Marchini. Haplotype estimation for biobank-scale data sets. *Nature genetics*, 48(7):817, 2016.

[10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[11] Ziad Al Bkhetan, Justin Zobel, Adam Kowalczyk, Karin Verspoor, and Benjamin Goudey. Exploring effective approaches for haplotype block phasing. *BMC bioinformatics*, 20(1):540, 2019.

[12] Yongwook Choi, Agnes P Chan, Ewen Kirkness, Amalio Telenti, and Nicholas J Schork. Comparison of phasing strategies for whole human genomes. *PLoS genetics*, 14(4):e1007308, 2018.

[13] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[14] Amy L Williams, Nick Patterson, Joseph Glessner, Hakon Hakonarson, and David Reich. Phasing of many thousands of genotyped samples. *The American Journal of Human Genetics*, 91(2):238–251, 2012.

[15] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[16] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.

[17] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

[18] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):7, 2015.

[19] Po-Ru Loh, Pier Francesco Palamara, and Alkes L Price. Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*, 48(7):811–816, 2016.

[20] Anthony Francis Herzig, Teresa Nutile, Marie-Claude Babron, Marina Ciullo, Céline Bellenguez, and Anne-Louise Leutenegger. Strategies for phasing and imputation in a population isolate. *Genetic epidemiology*, 2018.

[21] Shefali S Verma, Mariza De Andrade, Gerard Tromp, Helena Kuivaniemi, Elizabeth Pugh, Bahram Namjou-Khales, Shubhabrata Mukherjee, Gail P Jarvik, Leah C Kottyan, Amber

Burt, et al. Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in genetics*, 5:370, 2014.

[22] Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, Zhaohui S Qin, Heather M Munro, Gonçalo R Abecasis, et al. A comparison of phasing algorithms for trios and unrelated individuals. *The American Journal of Human Genetics*, 78(3):437–450, 2006.

[23] Cathy C Laurie, Kimberly F Doheny, Daniel B Mirel, Elizabeth W Pugh, Laura J Bierut, Tushar Bhangale, Frederick Boehm, Neil E Caporaso, Marilyn C Cornelis, Howard J Edenberg, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic epidemiology*, 34(6):591–602, 2010.

[24] Pierre Faux, Pierre Geurts, and Tom Druet. A random forests framework for modeling haplotypes as mosaics of reference haplotypes. *Frontiers in genetics*, 10:562, 2019.

## 4.7 Supplementary Methods

### 4.7.1 Dataset details

The data used in this study and not detailed in the main manuscript is described below:

1. **Reference dataset:** The reference panel for genotype imputation was formed purely from the HRC dataset by excluding 8,189 samples used in the Hap+HRCdataset and all samples from the 1000 genome project as they are from different populations.

2. **Low-density datasets**: Five different datasets were generated from chromosomes 16 and 21 of the Hap+HRCdatasets. PLINK was applied to remove a random SNP from each pair if the distance between them is less than the desired resolution. The used densities were 1 SNP per 1kb, 2kb, 3kb, 4kb and 5kb.

3. **Different population size datasets**: The datasets with different population sizes were generated by a random selection of the individuals from the Hap+HRCdataset. Five sample sizes were generated as follows: 500, 1000, 2000, 4000, 6000 and all individuals 8,258 for chromosome 16 and 21.

4. **Different ethnicity dataset**: This dataset was generated purely from HapMap III datasets for chromosome 21 and the same SNPs as the chromosome 21 in the Hap+HRCdataset mentioned above. Parents genotype data were excluded but used to resolve children's haplotypes for evaluation purposes. The final dataset contains 838 individuals as follows: African ancestry in Southwest USA (ASW): 43, Utah residents with Northern and Western European ancestry from the CEPH collection (CEU): 69, Han Chinese in Beijing, China (CHB): 84, Chinese in Metropolitan Denver, Colorado (CHD): 85, Gujarati Indians in Houston, Texas (GIH): 88, Japanese in Tokyo, Japan (JPT): 86, Luhya in Webuye, Kenya (LWK): 90, Mexican ancestry in Los Angeles, California (MXL): 27, Maasai in Kinyawa, Kenya (MKK ): 115, Toscani in Italia (TSI): 88, and Yoruba in Ibadan, Nigeria (YRI): 63.

Details about the datasets used in this study are summarised in Table 4.4. Genetic maps of the reference genome hg19 were obtained from public resources and used with the tools

for phasing and imputation when required. The HRC dataset is available through the European Genome-phenome Archive (dataset reference: EGAD00001002729). See Links below for data online resources.

**Table 4.4:** Details of datasets used in this study after preparation. Chromosomes' columns represent SNP count within each chromosome.

| Dataset | Samples | Chr2 | Chr6 | Chr11 | Chr16 | Chr21 |
|---|---|---|---|---|---|---|
| **Hap+HRC** | 8,258 | 50,521 | 38,930 | 30,580 | 21,928 | 9,319 |
| **Reference** | 16,481 | 3,392,237 | 2,460,111 | 1,936,990 | 1,281,297 | 531,276 |
| **HapMap-CEU** | 52 | 114,897 | 89,850 | 69,852 | 43,798 | 19,056 |

### 4.7.2 Consensus estimator construction

Consensus estimators of all possible combinations of three tools out of SHAPEIT2, SHAPEIT3, EAGLE2, and HAPI-UR were constructed and evaluated. We have shown previously that adding more tools does not improve accuracy if the added tool is substantially less accurate than the ones currently being used (See Chapter 3). Therefore, we limited the construction of our consensus estimator to only three different tools, shown to have the highest accuracies.

When constructing a consensus estimator from multiple iterations of a non-deterministic tool, we evaluated different combinations of the results. With respect to Hap+HRCdataset, the estimator constructed from all possible combinations of 3, 5, 7, 9 out of 10 different iterations of the same tool. For the consensus of a large number of iterations (50), we limited the results to 15 random combinations for each odd number from 3 to 49 out of 50.

We use a consistent naming convention to describe a particular implementation of consHap throughout this work. Individual tools are abbreviated ($S_2$: SHAPEIT2, $S_3$: SHAPEIT3, $E_2$: EAGLE2, $H_r$: HAPI-UR) and concatenated together if multiple tools are used. If multiple outputs from a single tool are used, we denote the number of iterations in parentheses. Thus consHap-$S_2E_2S_3$ is based on the output of SHAPEIT2, EAGLE2, and SHAPEIT3 while consHap-$S_2(3)$ is based on three different iterations of SHAPEIT2.

### 4.7.3 Evaluation of haplotype phasing

The haplotypes of 52 children from CEU cohort were resolved using their parental genotypes. Transmission phasing was able to resolve ∼75% of child heterozygous SNPs when at least one parent has a homozygous SNP in the same locus. We clarify that the parents' genotype data were excluded from the dataset used in this study as the focus is on phasing unrelated individuals. The same procedure was applied to all children from different populations to evaluate phasing accuracy on different ethnicity.

Phasing evaluation was conducted through switch error (SE) calculated for the 52 samples mentioned above similarly to these studies (Browning and Browning, 2011; Marchini et al., 2006). The average switch errors of the individuals are averaged and reported for the whole dataset. Missing SNPs, SNPs with Mendel errors, and unresolved SNPs (Child's SNP where both parents have heterozygous SNP within the same locus) were excluded from the evaluation.

### 4.7.4 Evaluation of genotype imputation

Hap+HRCdatasets were aligned to the reference panel datasets (The same alternate and reference allele for each common SNP between the reference and Hap+HRCdataset) which is mandatory for both Beagle5 and pbwt. Minimac3 seems to account for that as its results were not affected by the alignment. Imputation tools were applied to all phased chromosomes obtained from phasing experiments (by individual tools and consensus estimators). Imputation tools were applied using their default parameters. Genetic maps were only provided to Beagle5 as they are required for this tool. Reference panel was converted into pbwt format when provided to pbwt tool.

Imputation evaluation was carried out on 8,189 individuals from both UK10UK and IBD cohorts with respect to all SNPs in HRC dataset that do not belong to the SNPs of the Hap+HRCstudy (similarly to masked analysis concept (Verma et al., 2014)). Squared Pearson correlation $r^2$ as defined in Eq 4.1 was calculated for the real genotype and both the sum of the posterior allele probabilities (${r_a}^2$), and the imputed genotype (${r_g}^2$) for each SNP separately, similarly to the evaluations carried out previously (Browning et al., 2018; Das et al., 2016; Howie et al., 2012; Li et al., 2009). We clarify that $r^2$

is considered zero when the output of the imputation tool does not change for a single SNP across all individuals, as the correlation can not be determined in this case.

$$r_{(s,s')} = cor(s, s') = \frac{\sum_{i=1}^{n} (s_i - \bar{s})(s'_i - \bar{s'})}{\sqrt{\sum (s_i - \bar{s})^2 \sum (s'_i - \bar{s'})^2}} \tag{4.1}$$

Where: $s$ represents the value of the real genotype of a SNP as 0, 1, or 2. $s'$: represents the imputed genotype of a SNP in two possible forms: either a value of 0, 1, or 2 similar to the real genotypes representation when calculating ${r_g}^2$ or a value in the range (0, 2) as the sum of the posterior allele probabilities provided by the tools when calculating ${r_a}^2$. All evaluation metrics were averaged for all SNPs in three MAF ranges similar to the ranges in the study (Das et al., 2016): 0.01% to 0.5%, 0.5% to 5%, and 5% to 50%.

### 4.7.5 Correlation assessment of phasing and imputation accuracy

This analysis was carried out on 52 individuals from the HapMap project where their true haplotypes (resolved using family information) and true genotypes for the imputed SNPs are available. Both imputation and phasing accuracies were calculated at an individual scale for the five chromosomes. Imputation accuracy was measured as the percentage of the correctly imputed SNPs as a proportion of all individual's imputed SNPs. Only the common SNPs in HapMap III and HRC datasets that do not belong to the SNPs in the Hap+HRCdataset were assessed. The counts of the common SNPs are reported in the third row in Table 4.6. Phasing accuracy was calculated as (100 - Switch error %) also at an individual scale. We used a Pearson correlation to assess the relationship between each pair of phasing and imputation tools, with reported significance at a p-value threshold = 0.05.

## 4.8 Supplementary experiments

### 4.8.1 Evaluation at individual scale

Switch error results are averaged for switch errors calculated for the individuals in the dataset. While the consensus estimators reduced the overall switch error, there were some individuals in the dataset phased more accurately by one tool, or one iteration

of a non-deterministic tool compared to the consensus estimator. The switch error of consHap-$S_2E_2S_3$ and consHap-$S_2(3)$ for such cases was very similar to the most accurate tool or iteration (0.065% and 0.066% average difference, respectively) compared to the average difference between the most accurate two tools or iterations (0.2% and 0.1%, respectively). Table 4.5 shows the percentage of individuals that are phased by the tool in the related row compared to the tool in the related column. For example, 63.5 (row:1, column:1) means that 63.5% of all individuals were phased more accurately by the consensus consHap-$S_2E_2S_3$ compared to eh consensus consHap-$S_2(3)$. As we represent only the percentage of individuals that are phased more accurately by a tool compared to another one, the percentage of individuals that are phased similarly can be calculated from each two transposed cells. For example, the percentage of the individual phased with equal accuracy between consHap-$S_2E_2S_3$ and consHap-$S_2(3)$ is 100 - (Table2[consHap-$S_2E_2S_3$, consHap-$S_2(3)$] + Table2[consHap-$S_2(3)$, consHap-$S_2E_2S_3$]) = 100 - 63.5 - 31.2 = 0.3%.

Table 4.5 also highlights the performance of SHAPEIT2 and EAGLE2 individually, showing that 59% of CEU samples were phased more accurately by SHAPEIT2 compared to EAGLE2. HAPI-UR did not phase any individual more accurately than SHAPEIT2 or EAGLE2, but it did for 3% of the individuals compared to SHAPEIT3.

**Table 4.5:** Comparison of tool performance at an individual scale. Numbers represent the percentage of the individuals phased more accurately by the tool in the related row compared to the tool in the related column. Total number of individuals used to generate this table is 52 individuals for 5 chromosomes = 260 samples.

| – | **consHap-$S_2E_2S_3$** | **consHap-$S_2(3)$** | **SHAPEIT2** | **EAGLE2** | **SHAPEIT3** | **HAPI-UR** |
|---|---|---|---|---|---|---|
| **consHap-$S_2E_2S_3$** | 0.0 | 63.5 | 84.6 | 85.0 | 96.9 | 100.0 |
| **consHap-$S_2(3)$** | 31.2 | 0.0 | 75.4 | 73.1 | 93.5 | 100.0 |
| **SHAPEIT2** | 11.9 | 18.1 | 0.0 | 58.8 | 83.5 | 99.6 |
| **EAGLE2** | 12.3 | 23.1 | 39.2 | 0.0 | 75.8 | 99.6 |
| **SHAPEIT3** | 1.5 | 4.2 | 15.4 | 21.5 | 0.0 | 96.9 |
| **HAPI-UR** | 0.0 | 0.0 | 0.0 | 0.0 | 3.1 | 0.0 |

### 4.8.2 Performance evaluation

The performance evaluation was assessed based on the running time and memory usage during the execution of all phasing and imputation tools. We used a server of 32 VCPU, 2300 MHz and 64GB RAM operated by Ubuntu 16.04 LTS (Xenial). Tools applied with default parameters or specifying explicitly one thread for execution if this option is available. We don't report evaluation for all scenarios due to the need to do phasing

and imputation numerous times and in parallel using different thread count based on the available resources at a particular time.

Different manipulated versions of chromosome 21 dataset (the largest version was 19.1MB in plink format) were used for phasing evaluation. The same dataset phased by consHap-$S_2E_2S_3$ (23.6MB compressed VCF) in addition to chr21 reference panel (531,276 SNPs of 16,481 individuals, 534.5MB compressed VCF) were used to assess the performance of imputation tools.

The total time required to construct a consensus estimator is either the sum of the performance time for each tool when the tools are applied sequentially or the performance time of the slowest tool when the tools are applied in parallel in addition to the time required to aggregate phased haplotypes that can be negligible compared to the performance time of phasing tools. Constructing consHap-$S_2E_2S_3$ on chromosome 21 of the Hap+HRCdataset (large size and high density) required 32 hours. Phasing chromosome 21 of 500 individuals using SHAPEIT2 (the slowest tool in this study) required 40 minutes, therefore, the execution time of consHap-S2($n$) is $40 \times n$. However, these results represent the worst-case scenario where phasing is applied sequentially and using one thread.

The maximum RAM usage had a similar pattern to the performance time except for SHAPEIT3 that used more than SHAPEIT2. Phasing 9,319 SNPs of 8,258 individuals required less than 1 GB for HAPI-UR and EAGLE2, and 2 and 3 GB for SHAPEIT2 and SHAPEIT3 respectively.

Regarding genotype imputation performance, pbwt was the fastest tool accomplishing the imputation of the whole chromosome 21 within 40 minutes, 2 times faster than Beagle5 which required 77 minutes while Minimac3 spent 19 hours. pbwt was the optimal tool when considering memory usage. Beagle5 used ∼40GB of the ram during its execution. SHAPEIT2, SHAPEIT3 and Beagle5 can run using multiple threads, while Minimac3 requires another version. We have noticed after finalising the experiments that Minimac4 is available now, and it is claimed to be at least 2 times faster than Minimac3.

## 4.9 Supplementary tables

**Table 4.6:** A comparison of correctly phased haplotype blocks. BL: the average length of the correctly phased haplotype blocks at an individual scale. BC the average count of different haplotype blocks at an individual scale (equivalent to switch count). This table represents the reflection of the accuracy improvement obtained by the consensus estimator on increasing the length of the correctly phased haplotype blocks. The average length of correctly phased haplotype blocks obtained by consHap-$S_2E_2S_3$ and consHap-S2(3) is 421, and 403 SNPs respectively, representing improvements of 12% and 7.3% compared to the length of the correctly phased blocks obtained by the best individual tool in this test (∼40 SNPs longer than the most accurate individual tool). Best results are in bold. Best results obtained by an individual tool are underlined. The length of the correctly phased haplotype block was calculated at an individual scale as the SNP count between two consecutive switch errors occurred in heterozygous SNPs.

| **Phasing** | **Chr2** | | **Chr6** | | **Chr11** | | **Chr16** | | **Chr21** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **BL** | **BC** | **BL** | **BC** | **BL** | **BC** | **BL** | **BC** | **BL** | **BC** |
| **consHap-$S_2E_2S_3$** | **466** | **113** | **494** | **83** | **457** | **73** | **358** | **63** | **331** | **29** |
| **consHap-$S_2(3)$** | 450 | 116 | 470 | 85 | 439 | 76 | 356 | 65 | 302 | 31 |
| **SHAPEIT2** | 418 | 125 | 437 | 92 | 413 | 80 | 329 | 69 | 284 | 33 |
| **EAGLE2** | 398 | 129 | 424 | 94 | 392 | 83 | 314 | 71 | 287 | 34 |
| **SHAPEIT3** | 361 | 144 | 383 | 107 | 354 | 93 | 284 | 79 | 272 | 36 |
| **HAPI-UR** | 258 | 198 | 279 | 143 | 249 | 128 | 207 | 108 | 187 | 51 |

**Table 4.7:** Switch error % obtained by consensus estimator constructed from a combination of three tools (out of 4) calculated across 5 chromosomes. $S_2$: SHAPEIT2, $S_3$: SHAPEIT3, $E_3$: EAGLE2, and $H_r$: HAPI-UR. Consensus estimators are sorted based on the average switch error (lowest is first). Numbers in bold are the lowest switch errors.

| **Consensus Approach** | **Chr2** | **Chr6** | **Chr11** | **Chr16** | **Chr21** |
|---|---|---|---|---|---|
| **consHap-$S_2E_2S_3$** | **0.859** | **0.825** | **0.921** | **1.103** | **1.233** |
| **consHap-$S_2E_2H_r$** | 0.875 | 0.836 | 0.947 | 1.114 | 1.270 |
| **consHap-$S_2H_rS_3$** | 0.897 | 0.873 | 0.969 | 1.150 | 1.317 |
| **consHap-$S_3E_2H_r$** | 0.914 | 0.874 | 0.971 | 1.171 | 1.309 |

**Table 4.8:** Switch error % obtained by consensus estimator constructed from a combination of multiple iterations of different tools (out of 4) calculated across 5 chromosomes. S2: SHAPEIT2, S3: SHAPEIT3, and E2: EAGLE2. Numbers between parenthesis represent the number of iterations of the associated tool. Consensus estimators are sorted based on the average switch error (lowest error is first).

| **Consensus Approach** | **Chr16** | **Chr21** | **Mean-SE** |
|---|---|---|---|
| consHap-$S_2(8)S_3(0)E_2(7)$ | 0.992 | 1.122 | 1.057 |
| consHap-$S_2(7)S_3(0)E_2(6)$ | 0.995 | 1.131 | 1.063 |
| consHap-$S_2(6)S_3(0)E_2(5)$ | 1.002 | 1.135 | 1.068 |
| consHap-$S_2(5)S_3(0)E_2(4)$ | 1.010 | 1.135 | 1.073 |
| consHap-$S_2(4)S_3(0)E_2(3)$ | 1.015 | 1.141 | 1.078 |
| consHap-$S_2(7)S_3(0)E_2(8)$ | 1.010 | 1.155 | 1.082 |
| consHap-$S_2(6)S_3(0)E_2(7)$ | 1.018 | 1.162 | 1.090 |
| consHap-$S_2(5)S_3(0)E_2(6)$ | 1.020 | 1.173 | 1.097 |
| consHap-$S_2(4)S_3(0)E_2(5)$ | 1.032 | 1.175 | 1.104 |
| consHap-$S_2(7)S_3(7)E_2(7)$ | 1.038 | 1.174 | 1.106 |
| consHap-$S_2(3)S_3(0)E_2(2)$ | 1.035 | 1.185 | 1.110 |
| consHap-$S_2(5)S_3(5)E_2(5)$ | 1.045 | 1.176 | 1.111 |
| consHap-$S_2(3)S_3(0)E_2(4)$ | 1.041 | 1.184 | 1.112 |
| consHap-$S_2(3)S_3(3)E_2(3)$ | 1.047 | 1.178 | 1.112 |
| consHap-$S_2(2)S_3(0)E_2(3)$ | 1.050 | 1.198 | 1.124 |
| consHap-$S_2(1)S_3(0)E_2(2)$ | 1.090 | 1.231 | 1.161 |
| consHap-$S_2(2)S_3(0)E_2(1)$ | 1.104 | 1.224 | 1.164 |
| consHap-$S_2E_2S_3$ | 1.103 | 1.233 | 1.168 |

**Table 4.9:** Accuracy improvement when adding more iterations to the consensus construction. SE: the improvement in switch error. Switches: the reduction of switches. Numbers were calculated for the mean of switches, and switch errors for all combinations of the same number of iterations.

| **Iterations** | **Chromosome 16** | | **Chromosome 21** | |
|---|---|---|---|---|
| | **SE** | **Switches** | **SE** | **Switches** |
| **1 to 3** | 6.1% | 224 | 6.53% | 113 |
| **3 to 5** | 2.15% | 74 | 2.35% | 37 |
| **5 to 7** | 1.04% | 35 | 1.14% | 18 |
| **7 to 9** | 0.32% | 10 | 0.82% | 13 |

**Table 4.10:** ${r_a}^2$ and ${r_g}^2$ summarised for 37,442,564,700 SNPs with MAF in (0.0001, 0.005]. Without phasing, there were 37,322,838,335, 37,377,063,365, and 37,309,197,496 correctly imputed SNPs by Beagle5, Minimac3 and bwt, respectively.

| **Imputed by** | **Phased by** | $\boldsymbol{{r_a}^2}$ | $\boldsymbol{{r_g}^2}$ | **Accuracy** | **Improvement** |
|---|---|---|---|---|---|
| Beagle5 | consHap-$S_2(8)E_2(7)$ | 0.314 | 0.419 | 99.929 | 191,419 |
| Beagle5 | consHap-$S_2E_2S_3$ | 0.309 | 0.414 | 99.929 | 69934 |
| Beagle5 | consHap-$S_2(3)$ | 0.308 | 0.413 | 99.929 | 62303 |
| Beagle5 | EAGLE2 | 0.307 | 0.411 | 99.929 | 45799 |
| Beagle5 | SHAPEIT2 | 0.306 | 0.411 | 99.929 | 336900 |
| Beagle5 | SHAPEIT3 | 0.299 | 0.404 | 99.928 | 242223 |
| Beagle5 | HAPI-UR | 0.296 | 0.400 | 99.927 | 92611450 |
| Beagle5 | unphased | 0.066 | 0.137 | 99.688 | - |
| Minimac3 | consHap-$S_2(8)E_2(7)$ | 0.315 | 0.419 | 99.930 | 172308 |
| Minimac3 | consHap-$S_2E_2S_3$ | 0.310 | 0.414 | 99.930 | 66852 |
| Minimac3 | consHap-$S_2(3)$ | 0.309 | 0.413 | 99.929 | 39526 |
| Minimac3 | EAGLE2 | 0.308 | 0.412 | 99.929 | 52360 |
| Minimac3 | SHAPEIT2 | 0.307 | 0.411 | 99.929 | 304959 |
| Minimac3 | SHAPEIT3 | 0.299 | 0.404 | 99.928 | 194923 |
| Minimac3 | HAPI-UR | 0.297 | 0.400 | 99.928 | 38729491 |
| Minimac3 | unphased | 0.065 | 0.110 | 99.826 | - |
| pbwt | consHap-$S_2(8)E_2(7)$ | 0.279 | 0.382 | 99.922 | 242443 |
| pbwt | consHap-$S_2E_2S_3$ | 0.274 | 0.378 | 99.922 | 88760 |
| pbwt | consHap-$S_2(3)$ | 0.272 | 0.376 | 99.922 | 67560 |
| pbwt | EAGLE2 | 0.271 | 0.374 | 99.922 | 77275 |
| pbwt | SHAPEIT2 | 0.270 | 0.374 | 99.922 | 433473 |
| pbwt | SHAPEIT3 | 0.261 | 0.365 | 99.920 | 359676 |
| pbwt | HAPI-UR | 0.259 | 0.361 | 99.919 | 103500627 |
| pbwt | unphased | 0.047 | 0.100 | 99.654 | - |

**Table 4.11:** ${r_a}^2$ and ${r_g}^2$ summarised for 7,052,080,185 SNPs with MAF in (0.005,0.05]. Without phasing, there were 6,846,636,842 , 6,866,563,139, and 6,811,686,105 correctly imputed SNPs by Beagle5, Minimac3 and bwt, respectively.

| **Imputed by** | **Phased by** | $\boldsymbol{{r_a}^2}$ | $\boldsymbol{{r_g}^2}$ | **Accuracy** | **Improvement** |
|---|---|---|---|---|---|
| Beagle5 | consHap-$S_2(8)E_2(7)$ | 0.768 | 0.8589 | 99.395 | 502812 |
| Beagle5 | consHap-$S_2E_2S_3$ | 0.766 | 0.857 | 99.387 | 160331 |
| Beagle5 | consHap-$S_2(3)$ | 0.765 | 0.856 | 99.385 | 356462 |
| Beagle5 | SHAPEIT2 | 0.763 | 0.855 | 99.379 | 27218 |
| Beagle5 | EAGLE2 | 0.763 | 0.855 | 99.380 | 992514 |
| Beagle5 | SHAPEIT3 | 0.758 | 0.852 | 99.364 | 1635093 |
| Beagle5 | HAPI-UR | 0.751 | 0.847 | 99.341 | 160479842 |
| Beagle5 | unphased | 0.327 | 0.512 | 97.087 | - |
| Minimac3 | consHap-$S_2(8)E_2(7)$ | 0.764 | 0.855 | 99.386 | 504664 |
| Minimac3 | consHap-$S_2E_2S_3$ | 0.762 | 0.854 | 99.379 | 134018 |
| Minimac3 | consHap-$S_2(3)$ | 0.761 | 0.853 | 99.377 | 341527 |
| Minimac3 | SHAPEIT2 | 0.759 | 0.852 | 99.372 | 51177 |
| Minimac3 | EAGLE2 | 0.759 | 0.852 | 99.372 | 1000277 |
| Minimac3 | SHAPEIT3 | 0.754 | 0.849 | 99.356 | 1663102 |
| Minimac3 | HAPI-UR | 0.746 | 0.843 | 99.332 | 140092807 |
| Minimac3 | unphased | 0.298 | 0.454 | 97.369 | - |
| pbwt | consHap-$S_2(8)E_2(7)$ | 0.739 | 0.839 | 99.305 | 576712 |
| pbwt | consHap-$S_2E_2S_3$ | 0.736 | 0.837 | 99.296 | 207810 |
| pbwt | consHap-$S_2(3)$ | 0.735 | 0.836 | 99.293 | 368865 |
| pbwt | EAGLE2 | 0.734 | 0.835 | 99.289 | 43738 |
| pbwt | SHAPEIT2 | 0.733 | 0.835 | 99.287 | 1173304 |
| pbwt | SHAPEIT3 | 0.727 | 0.831 | 99.270 | 1983544 |
| pbwt | HAPI-UR | 0.719 | 0.824 | 99.241 | 188778197 |
| pbwt | unphased | 0.249 | 0.427 | 96.601 | - |

**Table 4.12:** ${r_a}^2$ and ${r_g}^2$ summarised for 9,735,689,186 SNPs with MAF in (0.05,0.5]. Without phasing, there were 8,881,462,946 , 8,857,189,955, and 8,805,437,448 correctly imputed SNPs by Beagle5, Minimac3 and bwt, respectively.

| **Imputed by** | **Phased by** | $\boldsymbol{r_a}^2$ | $\boldsymbol{r_g}^2$ | **Accuracy** | **Improvement** |
|---|---|---|---|---|---|
| Beagle5 | consHap-$S_2(8)E_2(7)$ | 0.971 | 0.983 | 99.218 | 256652 |
| Beagle5 | consHap-$S_2E_2S_3$ | 0.971 | 0.984 | 99.210 | 330655 |
| Beagle5 | consHap-$S_2(3)$ | 0.971 | 0.984 | 99.206 | 277583 |
| Beagle5 | EAGLE2 | 0.971 | 0.984 | 99.204 | 235839 |
| Beagle5 | SHAPEIT2 | 0.971 | 0.983 | 99.200 | 1460372 |
| Beagle5 | SHAPEIT3 | 0.970 | 0.983 | 99.185 | 1862852 |
| Beagle5 | HAPI-UR | 0.969 | 0.983 | 99.163 | 779068641 |
| Beagle5 | unphased | 0.745 | 0.853 | 91.177 | - |
| Minimac3 | consHap-$S_2(8)E_2(7)$ | 0.970 | 0.983 | 99.197 | 796271 |
| Minimac3 | consHap-$S_2E_2S_3$ | 0.970 | 0.983 | 99.188 | 344538 |
| Minimac3 | consHap-$S_2(3)$ | 0.970 | 0.983 | 99.184 | 227954 |
| Minimac3 | EAGLE2 | 0.970 | 0.983 | 99.183 | 284205 |
| Minimac3 | SHAPEIT2 | 0.970 | 0.983 | 99.178 | 1531159 |
| Minimac3 | SHAPEIT3 | 0.969 | 0.983 | 99.162 | 1698946 |
| Minimac3 | HAPI-UR | 0.969 | 0.982 | 99.143 | 801593644 |
| Minimac3 | unphased | 0.759 | 0.855 | 90.943 | - |
| pbwt | consHap-$S_2(8)E_2(7)$ | 0.968 | 0.982 | 99.139 | 863115 |
| pbwt | consHap-$S_2E_2S_3$ | 0.968 | 0.982 | 99.129 | 426841 |
| pbwt | consHap-$S_2(3)$ | 0.968 | 0.982 | 99.124 | 149534 |
| pbwt | EAGLE2 | 0.968 | 0.982 | 99.124 | 420757 |
| pbwt | SHAPEIT2 | 0.968 | 0.982 | 99.118 | 1674540 |
| pbwt | SHAPEIT3 | 0.967 | 0.981 | 99.100 | 1814525 |
| pbwt | HAPI-UR | 0.966 | 0.981 | 99.079 | 848101462 |
| pbwt | unphased | 0.725 | 0.837 | 90.425 | - |

**Table 4.13:** Correlation assessment for all combination of phasing and imputation tools averaged for chromosomes 21, 16, 11, 6, and 2. When evaluating genotype imputation accuracy on 8,189 HRC individuals (results are above and in the manuscript), we use the switch error calculated for only the 52 HapMap III individuals (with known real haplotypes) as an indication of the quality of pre-phasing for the whole dataset (HapMap III and HRC individuals combined in the Hap+HRCdataset). For a more accurate evaluation of the impact of phasing on the genotype imputation, we assess the correlation between phasing accuracy and genotype accuracy of the same individual. The lack of known haplotype data limited this experiments to 52 individuals from HapMap III dataset (Phasing evaluation can be conducted on them by comparing phased haplotypes to the one resolved using paternal genotypes).

| **Imputation tool** | **Phasing tool** | **Correlation** | **p-value** |
|---|---|---|---|
| Beagle5 | EAGLE2 | 0.422 | 1.93e-12 |
| Beagle5 | HAPI-UR | 0.486 | 1.58e-16 |
| Beagle5 | consHap-$S_2E_2S_3$ | 0.406 | 1.56e-11 |
| Beagle5 | SHAPEIT2 | 0.38 | 3.42e-10 |
| Beagle5 | consHap-$S_2(3)$ | 0.354 | 6.33e-9 |
| Beagle5 | SHAPEIT3 | 0.366 | 1.65e-9 |
| Minimac3 | EAGLE2 | 0.37 | 1.10e-9 |
| Minimac3 | HAPI-UR | 0.463 | 5.78e-15 |
| Minimac3 | consHap-$S_2E_2S_3$ | 0.34 | 2.48e-8 |
| Minimac3 | SHAPEIT2 | 0.317 | 2.43e-7 |
| Minimac3 | consHap-$S_2(3)$ | 0.285 | 3.64e-6 |
| Minimac3 | SHAPEIT3 | 0.316 | 2.52e-7 |
| pbwt | EAGLE2 | 0.31 | 4.29e-7 |
| pbwt | HAPI-UR | 0.399 | 3.75e-11 |
| pbwt | consHap-$S_2E_2S_3$ | 0.309 | 4.93e-7 |
| pbwt | SHAPEIT2 | 0.290 | 2.42e-6 |
| pbwt | consHap-$S_2(3)$ | 0.266 | 1.65e-5 |
| pbwt | SHAPEIT3 | 0.262 | 2.28e-5 |

## 4.10 Supplementary Figures

**Figure 4.5:** The relation between switch error% and the number of tool iterations used to construct a consensus estimator from a single tool. X-axis is the number of iterations. Results are summarised for all possible combinations of 3, 5, 7 and 9 combinations (out of 10 iterations) of each tool. We highlight that all these tools are non-deterministic by default but EAGLE2. Therefore, we changed the order of individuals within the dataset randomly with each iteration of EAGLE2.

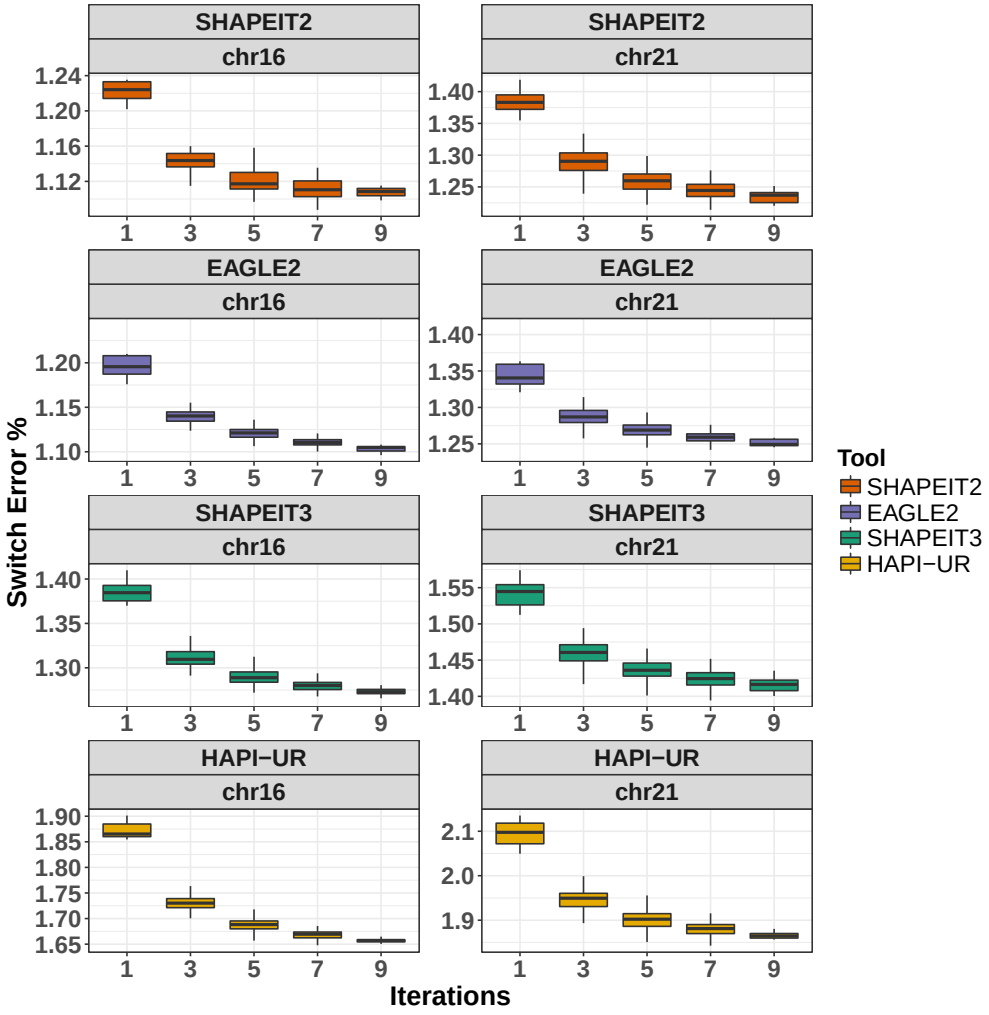


**Figure 4.6:** Comparison of the consensus approaches when applied to low and high-quality dataset from chromosome 16. a) Comparison of the performance of the consHap-$S_2E_2S_3$ (pink horizontal line) and a consensus approach of several iterations of SHAPEIT2 (in brown) (1, 3, 5, 7 and 9) applied to large and high SNP density dataset (chromosome 16 as described in Hap+HRCdataset in the manuscript). b) Comparison of the performance of the consHap-$S_2E_2S_3$ and a consensus approach of several iterations of SHAPEIT2 for all odd numbers between 1 and 50. The data used in b) is a low-quality version of chromosome 16 Hap+HRCstudy dataset, generated by reducing the sample size to 500 individuals and SNP density by 50%.

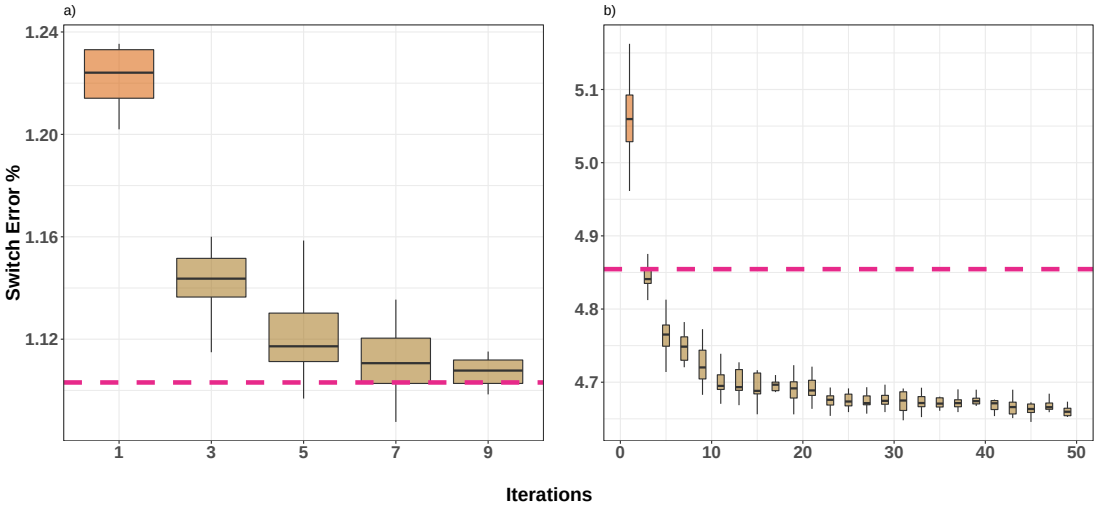


**Figure 4.7:** Switch error calculated for the consensus of SHAPEIT2, EAGLE2 and SHAPEIT3 (consHap-$S_2E_2S_3$) and the consensus of multiple iterations of SHAPEIT2 with respect to different population size and SNP density.

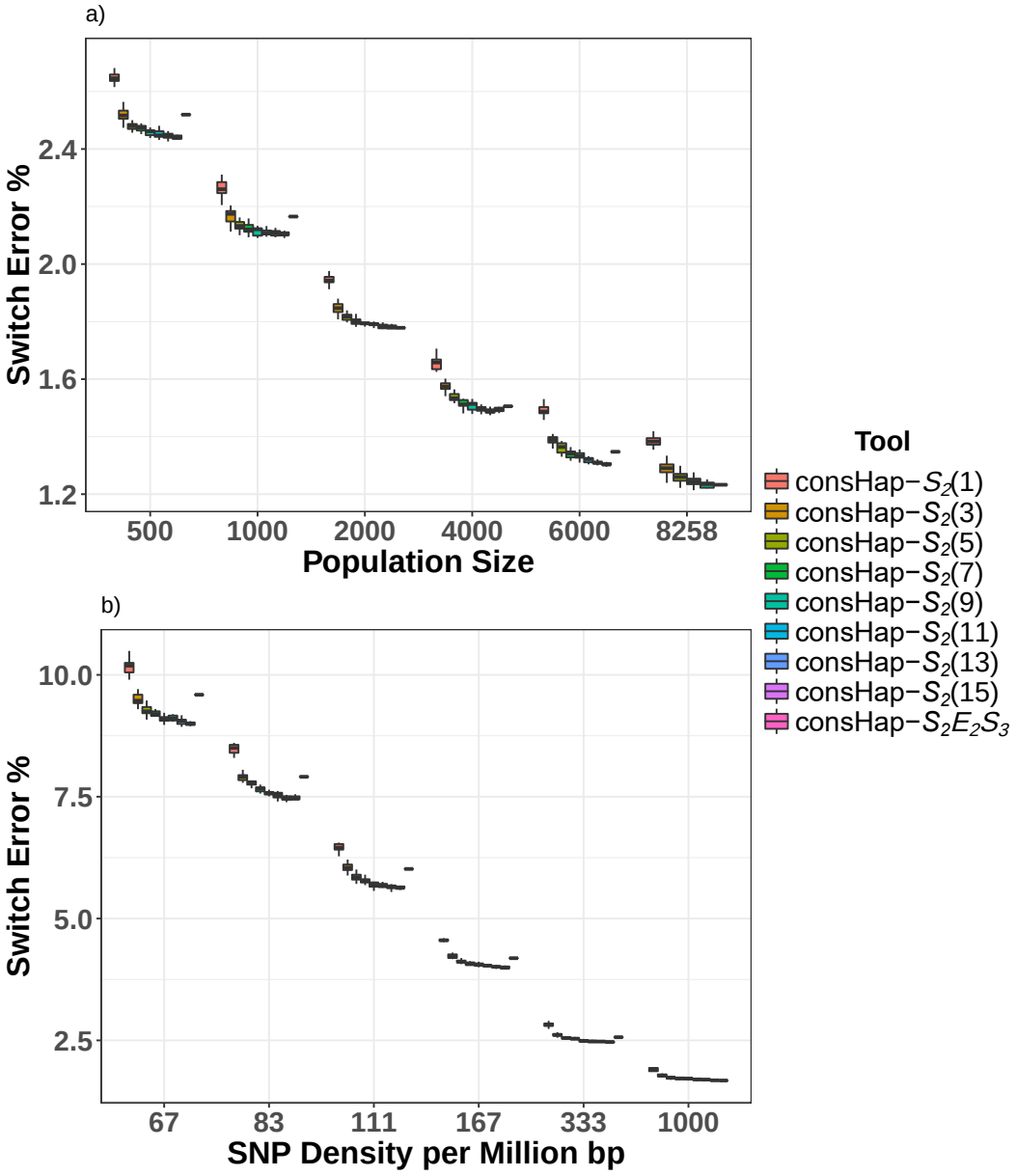


**Figure 4.8:** Comparison of the switch error obtained by consensus estimator approaches and individual tools when applied to the low-quality dataset. We clarify that both datasets contain only 500 individuals. SNP density in chromosome 16 was reduced to be 10,000 SNPs per the whole chromosomes. This figure shows that when EAGLE2 obtained a high error rate, the consensus consHap-$S_2E_2S_3$ performs similarly to three iterations of SHAPEIT2. Adding more iterations of SHAPEIT2 improves the accuracy more than what can be obtained by three different tools.

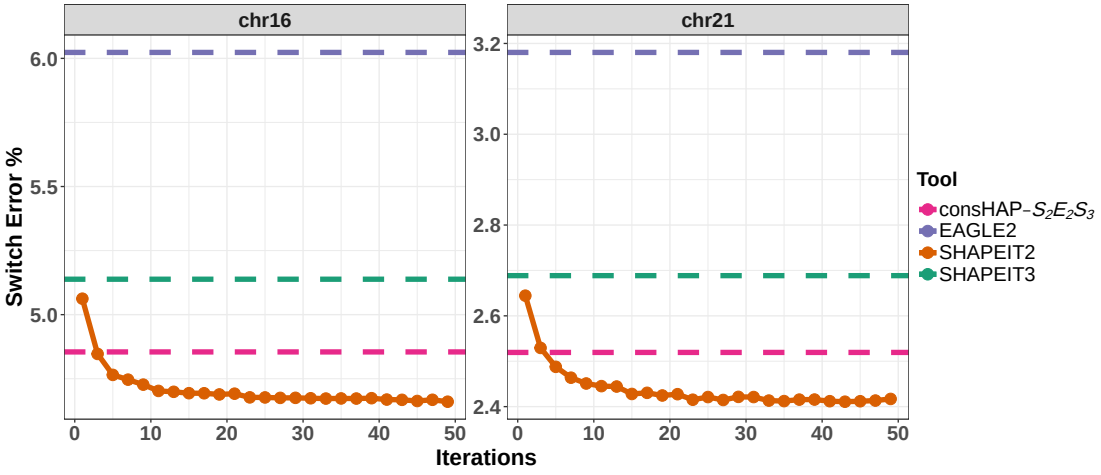



**Figure 4.9:** Comparison of the switch error obtained by a consensus estimator constructed by SHAPEIT2 and HAPI-UR for 1 to 49 iterations (only odd numbers). For each odd iteration ($i$), a 15 random combinations ($i$ iterations of 50 ) were used to construct a consensus of $i$ iterations. This figure shows that the accuracy of any consensus is highly associated with the tool used in its construction. 50 iterations of HAPI-UR doesn't outperform a single iteration of SHAPEIT2. Datasets used in this figure are described in the caption of Figure 4.8.

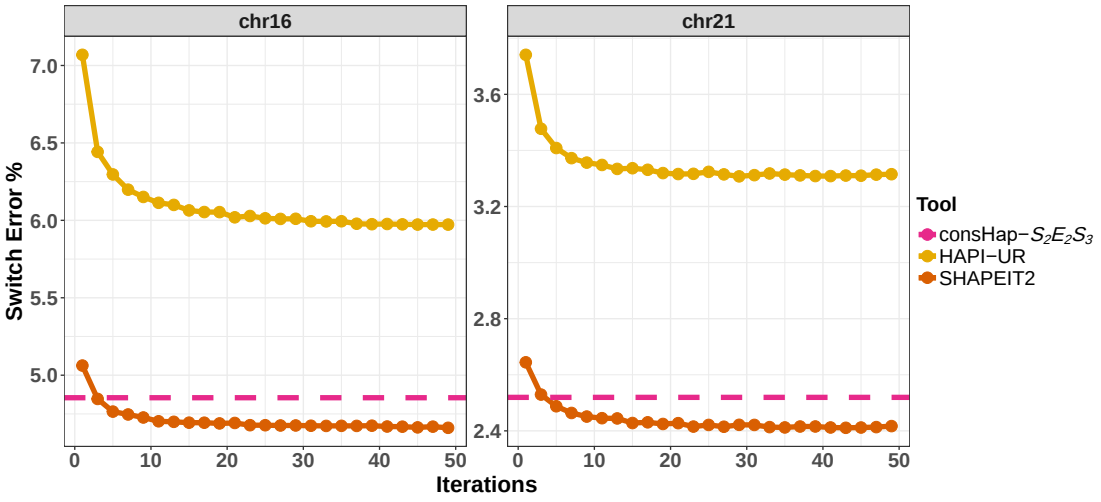


**Figure 4.10:** Correlation of phasing and imputation accuracy for 52 individuals from HapMap dataset. The x-axis represents phasing accuracy as 100 - Switch error %, while Y-axis represents the percentage of the correctly imputed SNP. Each point in the plots represents a pair of phasing and imputation accuracy for the same individual. The figure is generated from the results obtained from chromosomes 2, 6, 11, 16 and 21. The smoothed blue line represents a linear regression model of the imputation accuracy as a function of phasing accuracy. Each subplot is a combination of imputation and phasing tools. This figure is for the following phasing tools: consHap-$S_2E_2S_3$, SHAPEIT2, and EAGLE2.

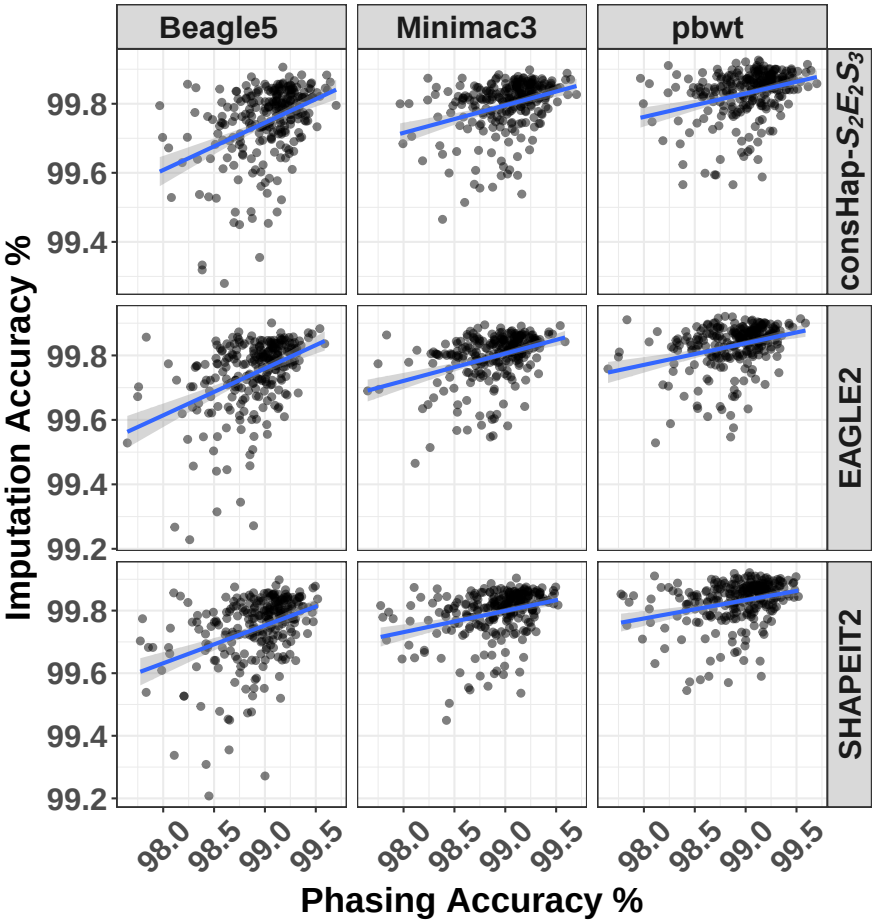
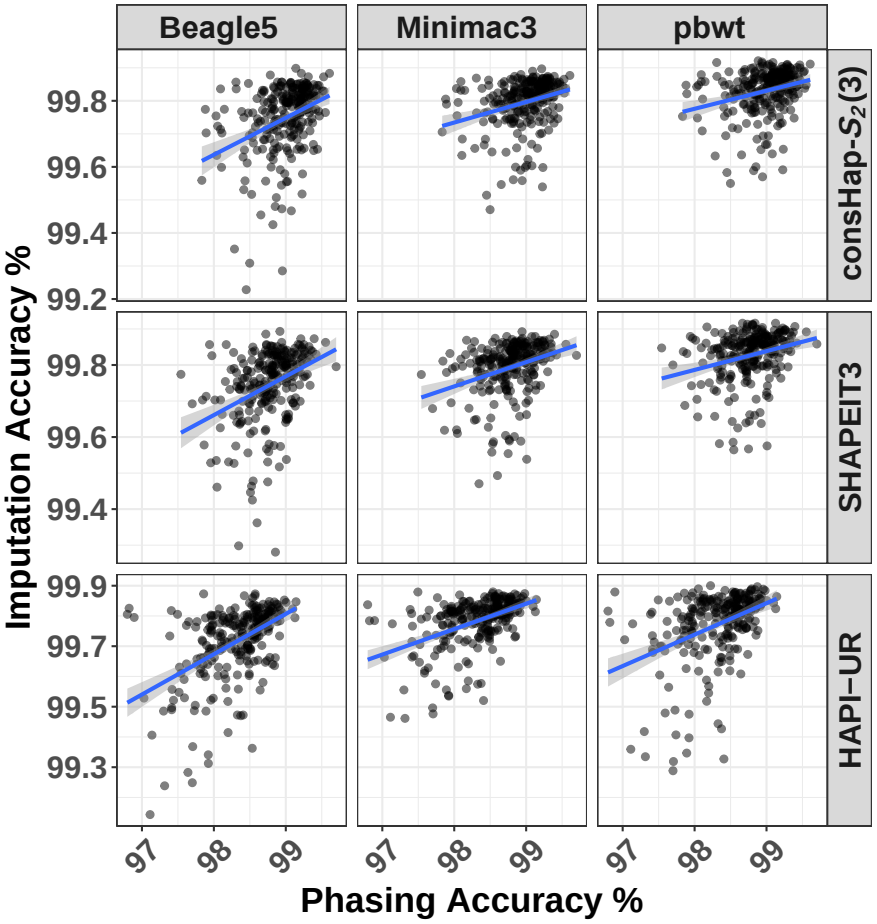


**Figure 4.11:** The same as Figure 6 but for the remaining phasing approaches: consHap-$S_2(3)$, SHAPEIT3 and HAPI-UR.



## 4.11 Links for online resource

1. **consHap tool**: https://github.com/ziadbkh/consHap.
2. **SHAPEIT2**: https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.
3. **SHAPEIT3**: https://jmarchini.org/shapeit3/.
4. **EAGLE2**: https://data.broadinstitute.org/alkesgroup/Eagle.
5. **HAPI-UR**: https://code.google.com/archive/p/hapi-ur.
6. **Beagle5**: https://faculty.washington.edu/browning/beagle/beagle.html.
7. **Minimac3**: https://genome.sph.umich.edu/wiki/Minimac3.
8. **PBWT**: https://github.com/richarddurbin/pbwt.
9. **Genetic maps 1**: https://data.broadinstitute.org/alkesgroup/Eagle/downloads.
10. **Genetic maps 2**: http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps.
11. **Haplotype Reference Consortium (HRC):** http://www.haplotype-reference-consortium.org.
12. **The HRC dataset:** Available through European Genome-phenome Archive (dataset reference: EGAD00001002729) at the link: https://www.ebi.ac.uk/ega/datasets/EGAD00001002729.
13. **HapMap III**: ftp://ftp.ncbi.nlm.nih.gov/HapMap/genotypes/HapMap3_r3.
14. **PLINK**: https://www.cog-genomics.org/plink2.
15. **Sanger imputation service**: https://imputation.sanger.ac.uk/.
16. **Michigan Imputation Server**: https://imputationserver.sph.umich.edu/index.html.

# Chapter 5

# eQTLHap: a tool for a comprehensive eQTL scan considering haplotypic and genotypic effects

*This chapter contains the investigations performed to respond to the following research question:*

> *Does including haplotype information in Expression Quantitative Trait Loci (eQTL)analysis lead to the detection of significant associations between genetic variations and gene expression that cannot be detected when analysing SNPs individually?*

This chapter appears in a manuscript[1] which is under review in Briefings in Bioinformatics - Oxford Academic. It is also available through bioRxiv at `https://www.biorxiv.org/content/10.1101/2020.07.23.206391v1.abstract`.

[1] Al Bkhetan, Ziad, et al. "eQTLHap: a tool for a comprehensive eQTL scan considering haplotypic and genotypic effects."

## 5.1 Motivation

The genetic contribution to diseases has been investigated by numerous studies at different molecular levels such as DNA (MacArthur et al., 2016) and RNA (Emilsson et al., 2008). Gene up/downregulation that can be associated with a disease is influenced by factors such as 1) Sequence information including cis- and trans-acting variations (Francesconi and Lehner, 2014; Tewhey et al., 2011), 2) Chromatin state including DNA-methylation and histone modifications (Bonifer and Cockerill, 2011), and 3) The three-dimensional structure of the genome that links promoters and enhancers located thousands or millions of base pairs away from their target gene (Andrey and Mundlos, 2017). In this chapter, we focus on one aspect of the three that is the sequence information.

The association between genetic variations (mainly single nucleotide polymorphisms SNPs) and gene expression has been investigated extensively on different datasets via Expression Quantitative Trait Loci (eQTL) studies. SNP-based eQTL studies assess statistically the relationship between the level of the gene expression and SNPs. SNPs can be assumed to have an additive or categorical effect on the gene expression (Shabalin, 2012). The assessment is applied for each gene and SNP pair where SNPs can be allocated on the gene, its surrounding regions (Consortium et al., 2017), regulatory elements related to the gene (Ying et al., 2018), or somewhere across the whole genome (Yap et al., 2018). These restrictions depend on whether cis or trans eQTL are targeted. In addition to the SNP-based analysis, studies investigated the association between gene expression levels and the epistatic interaction of SNP pairs (Brown et al., 2014; Hemani et al., 2014), structural variations such as insertion, deletion, and copy number variations (CNV) (Chiang et al., 2017). More details about eQTL analysis are available in the background, section 2.8 Expression quantitative trait loci (eQTL).

We explained in the background the dearth of haplotype-based investigation with respect to the eQTL problem. Including haplotype information increases the complexity of typical eQTL analysis for several reasons such as 1- Haplotype information is not readily available and most of the time it is computationally estimated which means extra time, effort and potential errors. 2- Genome partitioning into haplotype blocks is not a straightforward procedure as including extraneous SNPs or excluding important

ones can affect the results. 3- It is hard to identify an additive effect for a block of SNPs similarly to the case of one SNP which makes the encoding of haplotype blocks difficult. 4- Treating the haplotype block as a categorical effect is more natural but at the same time can lead to less statistical power when having multiple haplotypes within a block and few samples in the investigated dataset.

We utilise our findings in previous chapters regarding haplotype estimation evaluation and improvement in order to overcome the mentioned obstacles and to apply accurate haplotype-based eQTL analysis. More specifically, haplotype information is obtained via the consensus estimator that has been demonstrated to obtain not only accurate but also stable results more than any individual tool which helps to get replicable results.

In addition, haplotype blocks will be determined based on linkage disequilibrium (LD). LD-based blocks contain neighbouring alleles that exist together more often than being considered a random chance (Wall and Pritchard, 2003). Such blocks have been reported by GWAS to be associated with some diseases (Shang et al., 2015; Wu et al., 2014). We hypothesise that exploiting LD to determine haplotype blocks for haplotype-based eQTL analysis has the potential to detect accurate associations. Our reasoning is supported by the following:

1. Currently, haplotype information is obtained computationally via haplotype estimation "Phasing" tools (Browning and Browning, 2011). Phasing error rate within haplotype blocks determined by LD is substantially less than a sliding window approach as shown in Chapter 3.

2. There are very few different haplotypes within LD-blocks (Wall and Pritchard, 2003). This minimized diversity can lead to more statistical power as it reduces the degree of freedom when applying statistical assessment. This minimized diversity is more critical in eQTL analysis than GWAS due to the smaller sample size with eQTL analysis.

3. Haplotype blocks determined by LD is significantly less than blocks obtained by sliding window (especially overlapping sliding window). Having fewer blocks provides a less conservative significance threshold and more flexibility with multiple test correction.

4. There is evidence from GWAS studies that shows some associations between LD-blocks and specific diseases (Shang et al., 2015; Wu et al., 2014).

The final issue is related to haplotype encoding for association assessment. We propose bag-of-haplotypes encoding similar to the bag-of-words encoding used in text mining. This encoding helps to reduce the number of possible haplotypes within a block as it considers the paternal and maternal copies individually instead of combined.

In all previous studies, there was no comparison between the associations revealed by investigating single SNP and both the genotype/haplotype of the SNPs within blocks. Such evaluation is very essential to confirm whether including haplotype information that increases the complexity of the problem can lead to capturing novel associations or even detect the same associations but with higher significance compared with the genotypes within the same blocks.

# eQTLHap: a tool for comprehensive eQTL analysis considering haplotypic and genotypic effects

Ziad Al Bkhetan [1,2,*], Gursharan Chana [3], Cheng Soon Ong [2], Benjamin Goudey [1,4,†] and Kotagiri Ramamohanarao [1,†]

[1] School of Computing and Information Systems, The University of Melbourne, Victoria, Australia.

[2] Data61, CSIRO, Canberra, Australia.

[3] Department of Medicine, Royal Melbourne Hospital, The University of Melbourne, Victoria, Australia.

[4] IBM Research Australia, Victoria, Australia.

[*] ziad.albkhetan@gmail.com

## Abstract

**Motivation:** The high accuracy of current haplotype phasing tools has enabled the interrogation of haplotype (or phase) information more widely in genetic investigations. Including such information in eQTL analysis complements SNP-based approaches as it has the potential to detect associations that may otherwise be missed.

**Results:** We have developed a haplotype-based eQTL approach called *eQTLHap* to investigate associations between gene expression and haplotype blocks. Using simulations, we demonstrate that eQTLHap significantly outperforms typical SNP-based eQTL methods when the causal genetic architecture involves multiple SNPs. We show that phasing errors slightly impact the sensitivity of the proposed method ($< 4\%$). Finally, the application of eQTLHap to real GEUVADIS and GTEx datasets finds 22 associations that replicated in larger studies or other tissues and could not be detected using a single-SNP approach.

**Availability:** `https://github.com/ziadbkh/eQTLHap`.

[†]These authors have contributed jointly to this work as senior authors

# 1 Introduction

Genome-wide association studies (GWAS) have revealed numerous significant genetic associations with diseases. A large number of these variations can not be directly linked to a particular gene [1] as they are not part of the protein-coding regions [2]. Therefore, understanding how they influence a specific disease is a challenging task [3]. One mechanism underlying such associations is that genetic variations, especially those within regulatory regions, can affect gene transcription levels, reducing their functionality or deactivating them completely [4, 3]. As an example, around 50% of GWAS variations associated with schizophrenia have an impact on the expression of related genes [4]. Such cases are investigated statistically through expression quantitative trait loci (eQTL) analysis [5]. Numerous eGenes (genes whose regulation is influenced by SNPs) have been revealed through eQTL analysis applied to multiple populations and tissues [6, 3].

eQTL analysis has been widely applied using a range of approaches, primarily examining the relationship of single SNPs and gene expression [7]. In addition, joint analysis of multiple SNPs, interactions between SNP (epistasis) and phased haplotypes have also been considered [8, 9, 10]. Of these, haplotype-based approaches are amongst the least explored. However, it is well known that expression can be influenced by the phase of the mutations and the gene of interest. Such scenarios include compound heterozygosity, where a disorder is associated with two alternate alleles allocated on different homologous copies of a specific region, as well as allele-specific expression, where the allocation of mutations on each haplotype copy can have a different impact on the expression of homologous gene copies. Hence, phase-aware eQTL analysis is likely to have greater power than SNP approaches with respect to such cases [11].

A barrier for phased haplotype-based eQTL analysis is the increased complexity of the analysis. Many large eQTL studies rely on SNP array data and hence phase need to be estimated, a process which may introduce errors into the analysis. Moreover, decisions also need to be made about how to form haplotype blocks and how to represent these blocks when evaluating their associations with gene expression levels. We have shown that when dealing with haplotype blocks, the choice of partitioning method has an impact on the phasing errors [12]. Previous approaches using haplotypes for eQTL analysis have defined blocks from pairs of SNPs [9], combinations of up to 4 SNPs [10] or regulatory regions [13] but it is unclear how the accuracy of the determined haplotypes is influenced by these partitioning approaches. In contrast, recent work has shown that existing phasing tools can achieve high accuracy within the haplotype blocks defined using linkage disequilibrium (LD) [14, 12]. Such results encourage using LD-based blocks in this context. In addition, the use of LD will also minimise the diversity of the haplotypes within each block, leading to an increased statistical power to uncover associations with phenotypes of interest.

In this study, we present a method for haplotype-based eQTL analysis, called *eQTLHap*. Individuals' phased haplotypes are partitioned into variable-length blocks based on the LD between

SNPs. Using simulated genotype and gene expression data, we compare the detected significant associations when considering individuals' haplotypes to the ones detected using standard eQTL analysis (single SNP-based) and also considering the block genotype (combining all block's SNPs). The latter comparison demonstrates the importance of including the allele allocation in the analysis as both approaches consider the same combination of variants. Furthermore, we report the impact of phasing errors on eQTL results for different switch error rates (from 0% to 2.5%). Such novel analysis is essential to demonstrate the reliability and credibility of obtained results in real applications using available computationally phased haplotypes data. Finally, we applied our approach to two real genotype and gene expression datasets, GEUVADIS [3] and GTEx [6] and we investigate the revealed associations.

The results demonstrate the efficacy of the proposed approach for particular genetic architectures underlying the variation of gene expression. Considering phase information in eQTL analysis increases the true positive rate (TPR) of the detected eGenes, primarily when the causal genetic architecture involves multiple SNPs. The three eQTL approaches agree on a large percentage of the associations, yet each captures its own unique subset. There is a slight impact of phasing errors ($< 2.5\%$) on the TPR obtained by our method. eQTLHap uncovers associations (replicated in GTEx and GEUVADIS datasets) in blood that could not be detected by single SNPs but have been replicated in recent meta-analyses. These results highlight the value haplotype-based approached to complement current genotype-based methods for uncovering novel eQTLs.

# 2 Methods

## 2.1 Block determination and encoding

eQTLHap investigates haplotype blocks to conduct eQTL analysis. The blocks are encoded using three different ways depending on the conducted assessment SNP, block's genotype (denoted as B-Gen) or block's haplotype (denoted as B-Hap) as illustrated in Figure 1. SNPs are encoded considering an additive impact as the dosage of minor/reference allele as illustrated in Figure 1 b). This SNP encoding is similar to the standard SNP representation in most genetic problems and it does not account for the combined impact of multiple SNPs. The genotypic representation of a block is encoded by concatenating the genotypes of all SNPs within the block, and it is considered as a categorical variable as illustrated in Figure 1 c). This encoding accounts for the combined impact of multiple SNPs. Finally, the haplotypic representation of a block is represented using a bag-of-haplotypes encoding similar to the bag-of-words (BOW) encoding used in text mining [15]. The block of each individual is encoded as the dosage of each possible haplotype across all individuals as illustrated in Figure 1 d). This encoding accounts for both the combined impact of multiple SNPs as well as the allele location that is ignored by the genotypic encoding.

Haplotype blocks can be determined by several methods such as D-prime confidence interval [16], Four gamete [17], Solid spine [18], big-LD [19] and a simple sliding window. In this study, we used haplotype blocks determined based on LD though PLINK software [20] that implements

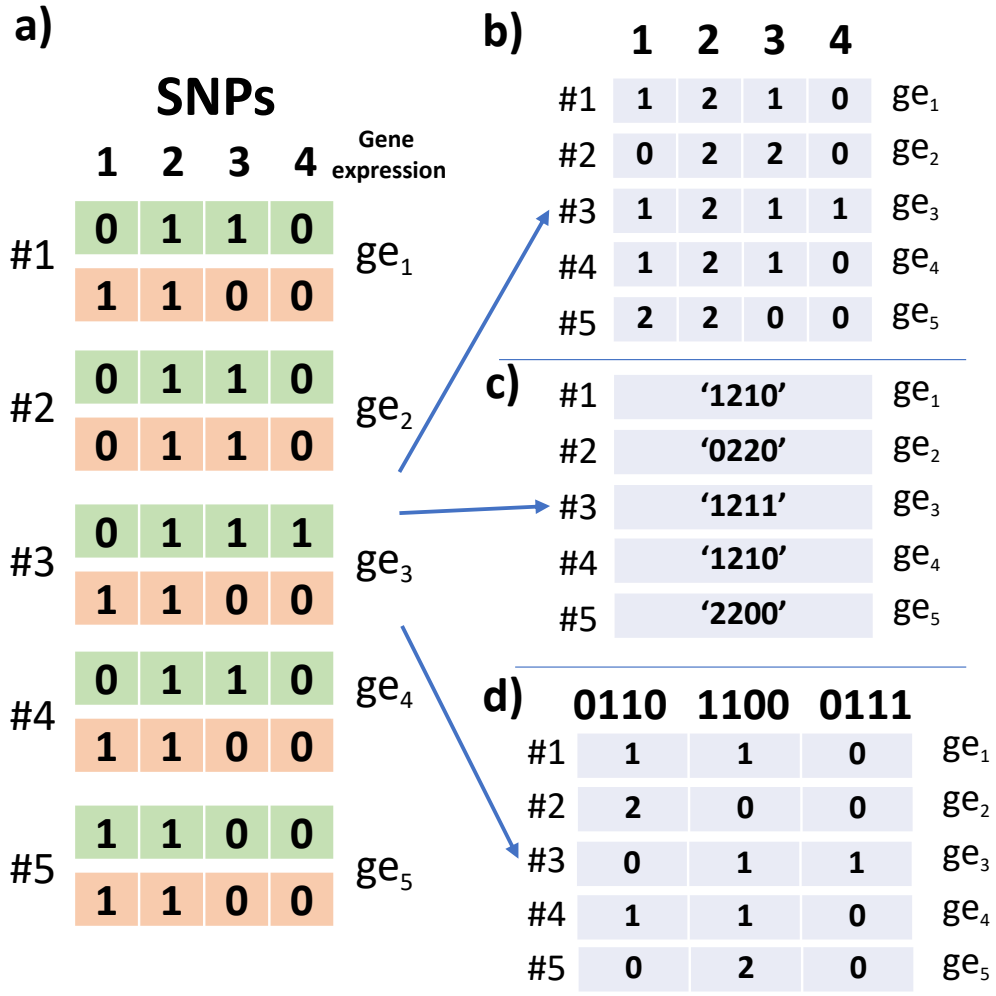



Figure 1: Different encoding for blocks. a) An example representing a block of 4 phased SNPs across 5 individuals and gene expression. Haplotypes are encoded using the dosage of the minor allele within the SNP (0: reference, 1: minor). b) SNPs are represented using their genotypes. c) B-Gen: Block's genotype is encoded by concatenating the genotypes of block's SNPs. d) B-Hap: Bag-of-haplotypes encoding for this block as the dosage of the three unique haplotypes within the block.

the Confidence Interval (CI) algorithm [16] (parameters: *–blocks no-pheno-req*). However, our eQTLHap accepts any kind of blocks (overlapping/non-overlapping, fixed or dynamic lengths) as long as the start and end SNP are determined for each block.

## 2.2 Association analysis

eQTLHap relies on applying eQTL analysis based on haplotype blocks and considering the alleles of both chromosome copies. After block determination and encoding as mentioned above, a multiple linear regression model is fitted for gene expression and the haplotype encoding, then $R^2$, F-test and p-value are calculated:

$$y = \sum_{i=1}^{m} \beta_i h_i + \sum_{j=1}^{c} \gamma_j v_j + \alpha; \quad \text{F-test} = \frac{(n-m-c)R^2}{m(1-R^2)} \tag{1}$$

where $y$ is the gene expression. $m$ is the number of unique haplotypes within the block (the columns in Figure 1 d). $h_i$ is the dosage of the haplotype $i$, $h \in (0, 1, 2)$. $v$ is a matrix of $c$ covariates. $n$ is the number of individuals. $c$ is the number of covariates. In addition to haplotype-based eQTL analysis (B-Hap), eQTLHap assesses the associations between gene expression and each SNP within a block (similar to the simple linear regression model provided by Matrix eQTL), as well as the genotype of the block (B-Gen). For block's genotype that is a categorical variable, an ANOVA regression model is fitted for this relation. The main difference between this ANOVA model and the model in Equation (1) is that here the genotype dosage is either 0 or 1. This comprehensive scan (SNP-based, B-Gen and B-Hap) was applied in all experiments reported in this study. The covariates part in Equation (1) is dropped out when such data is not included in the analysis.

SNPs with minor allele frequency (MAF) $< 0.01$ and haplotypes with frequency $< 0.02$ were eliminated to reduce the number of unique haplotypes per block. The unique haplotypes can be further reduced by considering haplotype tagging SNPs (htSNPs) as described in supplementary methods. However, tests on simulated data showed that complete haplotypes provide slightly better results as shown in supplementary figures 1 and 2, therefore, we confine the results in this manuscript to this configuration.

## 2.3 Implementation

eQTLHap is implemented in R and it depends on matrices operations to calculate correlation coefficients similar to the ultra-fast Matrix eQTL [7] to achieve high speed. It can be reconfigured to allow conducting block's haplotype, block's genotype and single SNP assessment individually or combined. In addition, it allows the adjustment of significance p-values through several methods accepted by *p.adjust* R function as well as configurable permutation analysis. Other parameters such

as covariate analysis, frequency thresholds can also be adjusted for customised analysis. Further implementation details are in the supplementary methods.

## 2.4 Genotype and haplotype preparation

Haplotypes are simulated using *msprime* simulator [21] using the same configuration to that of [22]. A region of 20 mbp was simulated for 1000 individuals (regular size in eQTL studies). Diallelic SNPs and SNPs with MAF $< 0.01$ were then eliminated.

GEUVADIS [3] and GTEx data from `https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/` and `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2` are used in our experiments. Quality control was applied to the genotype datasets using PLINK to keep only SNPs with MAF $> 0.01$, Hardy-Weinberg equilibrium (HWE) $< 10^{-6}$, the missing rate per individuals $< 0.1$, and missing rate per SNP $< 0.1$. Quality controlled genotype data were phased through *consHap* consensus phase estimator [23] by aggregating 15 applications of the SHAPEIT2 tool [24]. This approach has been reported to reduce the switch error rate significantly for datasets with small sample size ($<$ 2,000).

## 2.5 Gene expression simulation

Gene expression data were simulated for five causal genetic architectures similarly to the study [9] using the following equations:

$$\begin{aligned} Null: \quad & y = \epsilon \quad where \quad \epsilon \sim \mathcal{N}(0,1) \\ Causal: \quad & y = \beta + \epsilon \quad where \quad \epsilon \sim \mathcal{N}(0, var(\beta)\frac{1-\sigma^2}{\sigma^2}), \end{aligned} \tag{2}$$

where $y$ is the simulated gene expression, $\beta$ is the assumed effect size of the underlying variant (single SNPs, pairs of SNPs or haplotypes) while $\sigma$ is the gene expression heritability, i.e the proportion of the expression variation caused by the genetic architecture. We consider models where causal variant were common (MAF $\geq 0.05$) or rare (MAF $< 0.05$) SNPs ($\beta = g$, where $g$ is minor allele dosage of a SNP as illustrated in Figure 1 b), common and rare haplotypes ($\beta = h$, where $h$ is the dosage of the haplotype) together with pairs of SNPs (additive where $\beta = g_1 + g_2$ and interaction where $\beta = g_1 \times g_2$). Pairs of rare, common and a mix of rare and common SNPs were considered. In addition, we simulated expression under a null model with no variants having an effect to identify false-positive rates. $\sigma$ was fixed at 0.05 in all simulations, with the impact of changing $\sigma$ reported in Supplementary Figures 3-5.

For simulation based on a pair of SNPs, pairs within 7.5 kbp were picked (within blocks and randomly) when their correlation (squared Pearson correlation) is $< 0.8$ and none of them has a correlation $> 0.8$ with the encoding of their interaction or additive impact. Similarly, causal

haplotypes were picked when they do not have a correlation $> 0.8$ with any single SNP within the same block. This limitation avoids the cases where these simulations will be equivalent to single SNP-based simulations. For each experiment, 100 simulations were run for each causal architecture with random initialisation. Further details of these simulations are provided within the supplementary methods.

## 2.6 Simulations to compare different eQTL approaches

To reduce the impact of synthetic data on the results, we used real genotypes for the simulations in this experiment. Genotype data were obtained from the 1000 Genome Project for 373 individuals from European population (obtained from GEUVADIS dataset). Genotypes of chromosome one were phased using SHAPEIT2 [24] and haplotype blocks were determined using PLINK software. Five genomic regions were extracted for these individuals within 1mbp up/downstream of the following arbitrarily chosen genes ENSG00000218510, ENSG00000127074, ENSG00000196539, ENSG00000162441, and ENSG00000198468. For each of the give genes, we simulate 100 common and 100 rare SNPs, leading to 1000 simulations for the single SNP and haplotype architectures. For pair based architectures, we simulate 100 pairs of common SNPs, 100 pairs of rare SNPs and 100 rare/common pairs for the give genes, and repeating simulations to consider pairings within a single block and pairings between blocks, leading to 6000 simulations in total. As such, 14,000 gene expression simulations were generated for all mentioned casual architectures. eQTL analysis is performed using *eQTLHap* for all representations (SNP, B-Gen and B-Hap).

## 2.7 Impact of phasing error on haplotype-based eQTL

To assess the impact of phasing accuracy on eQTL results, we generated five haplotype datasets from the same region with different switch errors (SE) ranging from 0 to 2.5%. We conducted eQTL analysis for each dataset using the same simulated gene expression. For this experiment, a genotype dataset was formed from simulated haplotypes by combining both haplotype copies of individuals. To obtain realistic SEs, the formed genotypes were then phased using HAPI-UR [25] 100 times and all SEs were recorded. HAPI-UR was used as it is fast and non-deterministic by default [12]. The average SE of the 100 applications was 0.78%, while the maximum SE obtained when considering all unique SE locations was 2%. Six versions of the simulated region were prepared by switching individual's haplotypes within the locations recorded for HAPI-UR's SEs. SE within these datasets varies from 0% (the original simulated data with no errors) to 2.5% by 0.5% step. For the version with SE = 2.5%, in addition to all recorded SE (account for 2%), we used random heterozygous SNPs as locations of SEs to reach the desired SE (2.5%).

Four different regions (1.5 mbp) were selected arbitrarily from the simulated haplotypes (SE = 0%). Haplotype blocks determined using PLINK software and overlapping with each region are

used for further analysis. Gene expression data were generated as explained above. Haplotype-based eQTL analysis (B-Hap) was applied for the simulated gene expression and the 6 versions of the haplotypes (different switch errors) for each region.

## 2.8 Analysis of GEUVADIS and GTEx data

The comprehensive eQTL analysis was applied to the GEUVADIS dataset following similar configurations as its originally reported [3]. Briefly, individuals of European and Yoruba populations (373 and 89, respectively) were analysed separately. Gene expressions were used after probabilistic estimation of expression residuals (PEER) normalisation [26]. The top three principal components (PC) were used as covariates within individuals of European descent, while the top 2 PCs were used for Yoruba individuals to eliminate any impact of population stratification.

GTEx gene expression and covariates for multiple tissues were obtained from the GTEx web portal `https://www.gtexportal.org/home/datasets`. GTEx's covariates datasets include hidden factors detected by PEER normalisation. For both datasets, genes with non-zero expression level for more than 90% of the individuals were investigated considering SNPs within 1mbp up/down from their transcription start site (TSS). Associations with "empirical" p-value $< 0.05$ were recorded for multiple test correction based on a permutation of 1,000 iterations.

## 2.9 Evaluation and comparison

After conducting eQTL analysis, p-values are reported for all associations. Multiple test correction (MTC) was carried out for all p-values (SNPs, block's genotype, and block's haplotypes, separately) using the Benjamini-Hochberg (BH). Associations based on simulated data are considered significant when their BH corrected p-value is $< 0.05$. Associations based on real data are considered significant when their BH corrected p-value is $< 0.05$ and the permutation-based p-value is $< 0.015$.

With simulated gene expression data, TPR was calculated as the percentage of detected simulated associations of all simulated associations with respect to each causal architecture. For a model to have 100% TPR for simulations based on SNP pairs, both pairing SNPs should be reported as significant. If only one SNP of each pair was detected as significant for 100 different simulations, the TPR will be 50%. For haplotype-based eQTL, the block containing the causal SNP should be reported significant. When applying SNP-based eQTL on haplotype-based simulations, if any SNP within the causal haplotype block is reported significant, the association considered detected. Venn diagrams were generated for each causal architecture, where simulations for pairs account for 2 causal SNPs.

Significant blocks reported for GTEx were transformed from GHR38 to hg19/GHR37 using the LiftOver web tool `https://genome.ucsc.edu/cgi-bin/hgLiftOver` to match the same genome assembly as GEUVADIS data. After that, replications between GTEx and GEUVADIS eGenes are

identified when a significant block from GTEx overlaps with another one from GEUVADIS for the same gene.

# 3 Results

## 3.1 Comparison of different approaches for eQTL analysis

We compared the results of eQTL analysis using single SNPs, SNP blocks as haplotypes (denoted B-Hap) and as genotypes (denoted B-Gen) using genotypes from 273 European individuals from the 1000 Genome Project and simulated gene expression under five causal architectures. Within 5 genes, the average number of haplotype blocks is 816 with an average length of 1.25 kbp and an average SNP count of 17.

Figure 2 shows that B-Hap analysis is superior to other approaches when the causal architecture involves a haplotype stretch, an additive impact of a SNP pair or an interaction of rare SNPs. As expected, simulation-based on single SNPs are detected better via SNP-based eQTL regardless of the frequency of the SNP. B-Gen approach was more effective when dealing with interactions of common SNPs. With respect to the null causal architecture, the FDR was 0.06%, 0.05% and 0.05% for SNP, B-Hap, and B-Gen, respectively. While these results are reported for $\sigma = 0.05$, a similar pattern was observed when changing $\sigma$ from 0.01 to 0.1 as shown within supplementary Figures 3, 4, and 5. The main difference is that the detection rate increased for all approaches with $\sigma$.

We further compared the simulated associations detected by the three approaches to quantifying the similarity and differences between these models. The Venn diagram shown in Figure 3 illustrates that each approach could identify a unique set of the simulated associations that was not captured by others, with B-Gen being the least effective method. With a SNP-based causal architecture, a SNP-based eQTL analysis has the highest power to detect associations, detecting the casual variant in 980 out of 1000 simulations. The significance of haplotype-based analysis is demonstrated with other causal architectures where this approach could reveal a large number of associations that were ignored by SNP-based analysis. We have observed from detailed results, that SNP-based analysis could detect one SNP of causal pairs for the majority of the simulations, yet, the other pairing SNPs were missing. With such cases, the haplotype-based analysis could capture both SNPs involved in the simulations. As expected, we observe a drop in TPR for all approaches when comparing results from real vs simulated genotypes.

These experiments show that the three approaches not only agree on a large percentage of the detected associations but also complement each other by revealing a unique subset of the simulated associations.

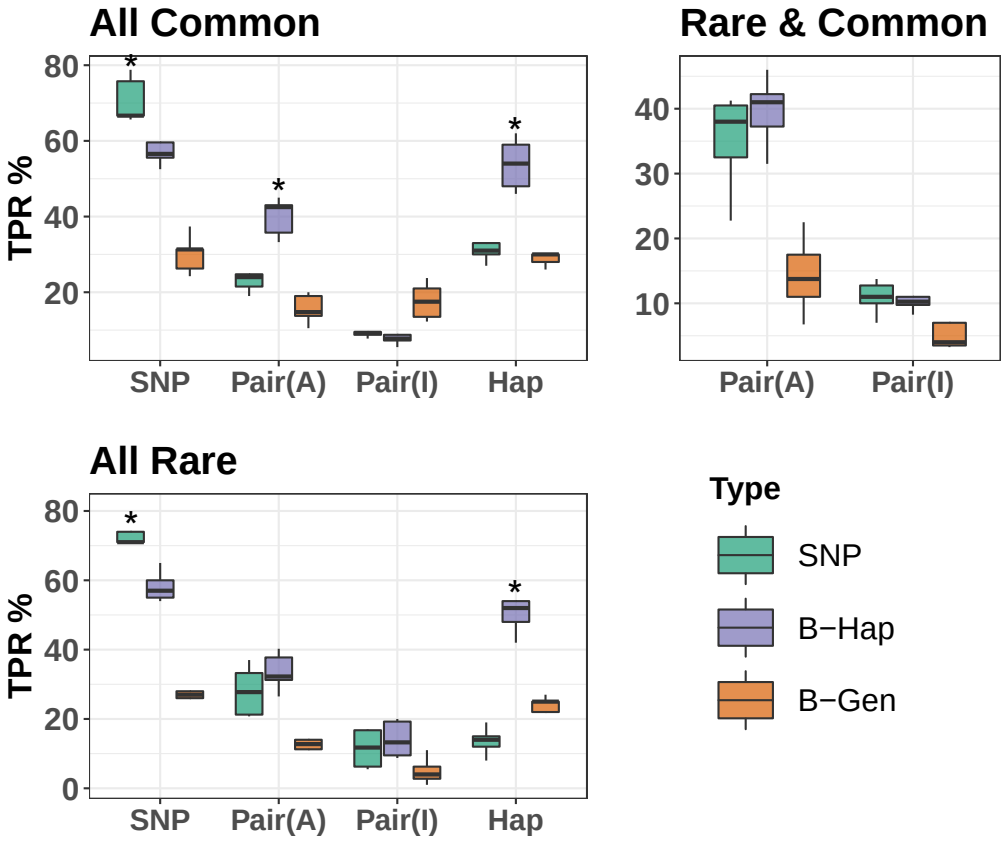



Figure 2: TPR for eQTL analysis based on SNP, B-Hap, B-Gen when applied to simulated genotype and gene expression data for different causal architectures. The x-axis represents the causal architecture where Pair(A) is an additive impact of a SNP pair and Pair(I) is an interaction of a SNP pair. The 'Rare & Common' scenario is only available for SNP pairs. Asterisks represent results that are significantly higher than all other approaches (using t-test and significance threshold of 0.05).

### 3.2 Impact of phasing errors on haplotype-based eQTL

In real applications, haplotype information is not perfect as they are obtained computationally through phasing methods. Therefore, it is important to assess the impact of the errors on the downstream haplotype-based eQTL analysis. Here, we assess the impact of SE, the standard metric of phasing evaluation [14], on the sensitivity or TPR of haplotype-based eQTL analysis applied to four regions of 1.5 mbp (1,945 SNPs) using simulated genotype and gene expression data.

Haplotype-based eQTL analysis was applied to an average of 108 haplotype blocks within each

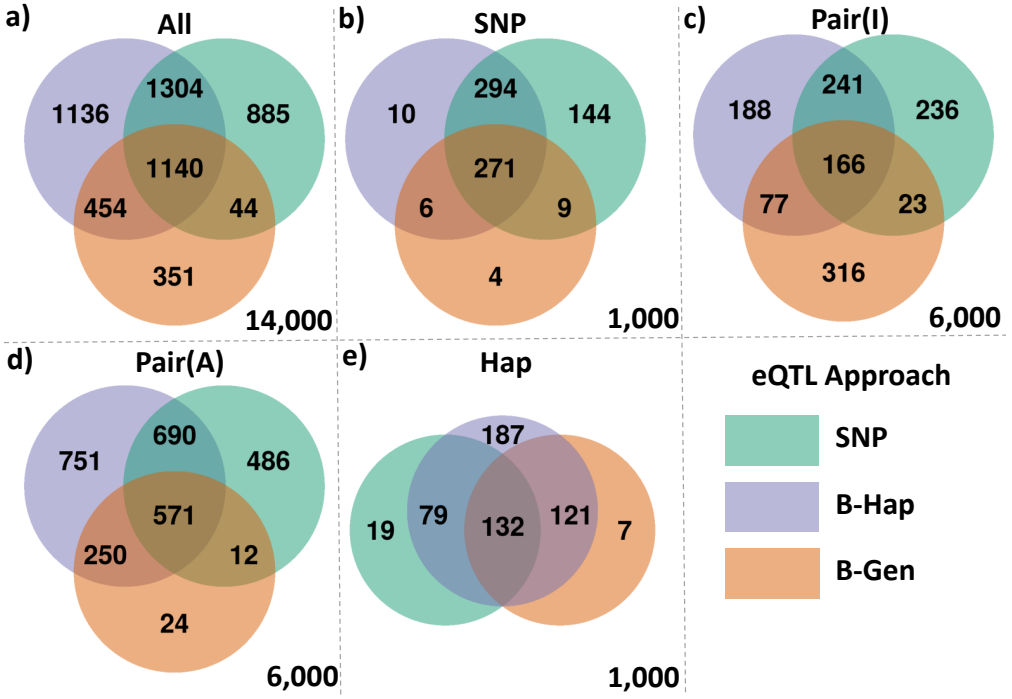



Figure 3: Venn diagram for significant associations detected by B-Hap, B-Gen and SNP based eQTL on simulated data. Subplots correspond to different causal architectures: a) all simulations combined. b) SNP. c) interaction of SNP pair. d) additive effect of SNP pair. e) haplotype. The total number of simulated associations are in right bottom corners.

region. Block lengths varied from 0.03 kbp to 59 kbp with an average length equals to 12 kbp (mean SNP count is 16, min = 2 and max = 68). 6,400 simulations were generated for all causal architectures followed by association assessment for each block-gene expression pair. This analysis was repeated for 6 versions of the four regions where SE varied from 0 to 2.5% by 0.5% step. The percentage of incorrectly phased haplotype blocks within these datasets were 0.4% 0.8%, 1.1%, 1.7%, and 3.2%, respectively.

We observed that the number of reported significant associations reduced by an average of 1,495 associations when SE increased from 0 to 2.5%. After MTC, the false discovery rate (FDR) when there is no genetic causal was between 0.04% and 0.06%. Figure 4 a) shows the percentage of detected simulated associations varied from 7% to 100% depending on the causal architecture and phasing error within the data. There was a slight impact of SE within the range (0-2.5%) on the TPR of haplotype-eQTL analysis, especially when the causal architecture involves a common SNP/haplotype (solid lines in the figure). There was 3.9%, 3.6%, 2%, and 1.8% TPR reduction when SE increased from 0 to 2.5% for single SNP, an additive impact of SNP pair causal architectures,

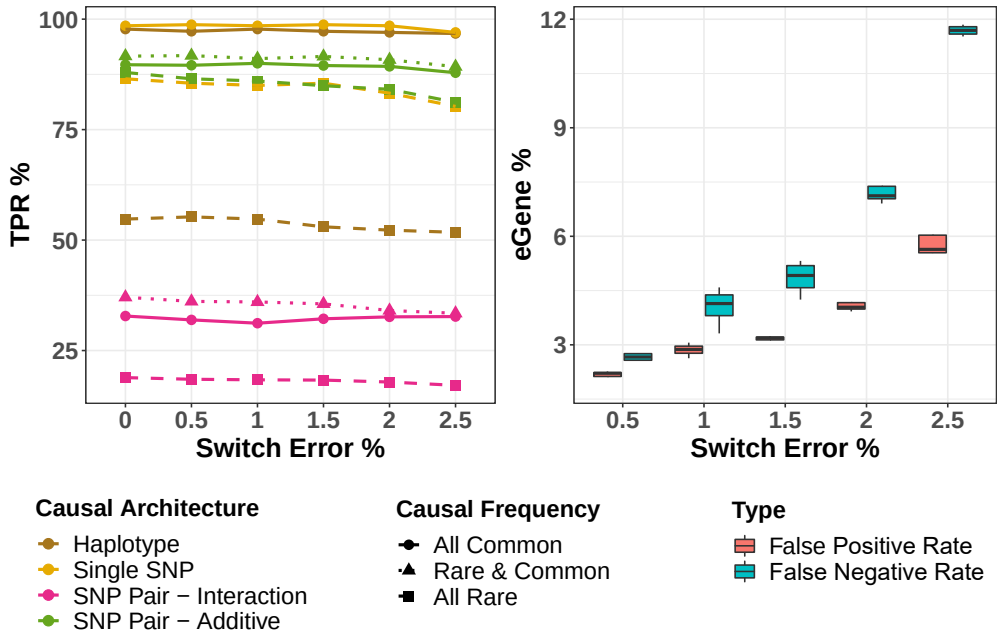



Figure 4: Impact of SE on haplotype-based eQTL analysis. a) TPR of the analysis with respect to switch error. b) Percentage of false positives/negatives with respect to SE.

haplotype, and an interaction of a SNP pair respectively. SNP or haplotype-based association were easier to detect compared to the associations based on SNP pairs. We observed a low detection rate of simulations based on interactions of SNP pairs (less than 37%) compared to other casual architectures.

Furthermore, we investigated how the detected significant associations vary with respect to different SEs compared to the associations revealed when there were no errors present (i.e. SE = 0). Haplotype-based eQTL conducted on the data with SE = 0 reported 43,249 significant associations with a corrected p-value $< 0.05$. These associations represent 3% of all block-gene expression combinations tested. 11% of these associations are the true simulated associations. The comparison of these associations and the ones revealed from datasets with different SEs, shown in Figure 4 b), shows that when SE increases, the percentage of false positives (detected with SE $> 0$ dataset but were not detected when SE = 0) and false negatives (detected when SE = 0, but not detected when SE $> 0$) increases reaching 5.9% and 11.7%, respectively. However, the true simulated associations seem to be more robust against SE as the TPR illustrated in Figure 4 a) is less affected. Investigating the false positives showed that few of them were from the simulated associations (reaching 2.55%) but not captured when SE equals 0. Around 4.5% of the

false negatives were from the simulated associations. 11.3% of the true positives (shared between dataset with SE = 0 and SE > 0) are true simulated associations.

These results support the application of haplotype-based eQTL analysis on real datasets where SE is usually less than 2.5% [12] especially given there is little impact on the TPR.

### 3.3 Application to GEUVADIS and GTEx data

After we demonstrated the efficacy of the approach through synthetic data, we applied the comprehensive eQTL analysis to real genotype/haplotype and gene expression dataset from both projects GEUVADIS (lymphoblastoid cell line for European and African populations) and GTEx (whole blood, artery coronary, and brain amygdala tissues).

In both GEUVADIS and GTEx, the overlap of eQTLs discovered by each representation, shown in Figure 5, shows a similar pattern to those in the simulated data. The approaches agree on a large set of associations with each approach finding a subset of unique associations. Similar trends can be seen in other tissues types (Supplementary Figure 6).

Applying eQTLHap to the whole blood tissue of GEUVADIS (EUR population) and GTEx, there were 14,229 (of 76,146) and 36,696 (of 187,644) significant haplotype-based associations whose p-values are less than 100 times the p-value of both block's genotype and each SNP within the same block. These results include 1,035 and 748 genes that only haplotype-based eQTL could detect significant associations with. The average p-value of these eGenes is $2\times10^{-4}$ and $2\times10^{-4}$ within GTEx and GEUVADIS, respectively.

Figure 6 shows an example of a haplotype-only association, plotting the distribution of expression for *USP46-AS1* in whole blood from the European population of GEUVADIS across different variant representations for a block of three intronic SNPs (rs7657404, rs7698053, rs7688816). In this instance, the haplotype analysis is highly significant (q-value$<$ 0.017), while the single SNP and genotype block representations are not (q-values 0.9 and 0.99, respectively). Simplifying the haplotype analysis in Figure 6 b), we see that the individuals carrying the CTG haplotype have substantially higher gene expression than those who do not. These three SNPs were found in GTEx to be significant eQTLs for RASL11B in Testis and DANCR in cultured fibroblasts [27] but we believe this to be the first association with *USP46-AS1* in whole blood reported to date.

Furthermore, we searched for replicable, significant associations from our haplotype-based eQTL analysis from GTEx and GEUVADIS dataset for whole blood tissue. There were 50,616 common associations for 2,425 unique genes reported for both datasets with 3,136 common genes between both datasets. The average p-value of these common associations is $4\times10^{-4}$ and $9\times10^{-4}$ with respect to both GTEx and GEUVADIS, respectively. From these 2,425 genes, there were 11 common eQTLs for 7 eGenes (Supplementary Table X) that only had significant associations when they were represented using B-Hap. To further validate these findings, we compared these associations with

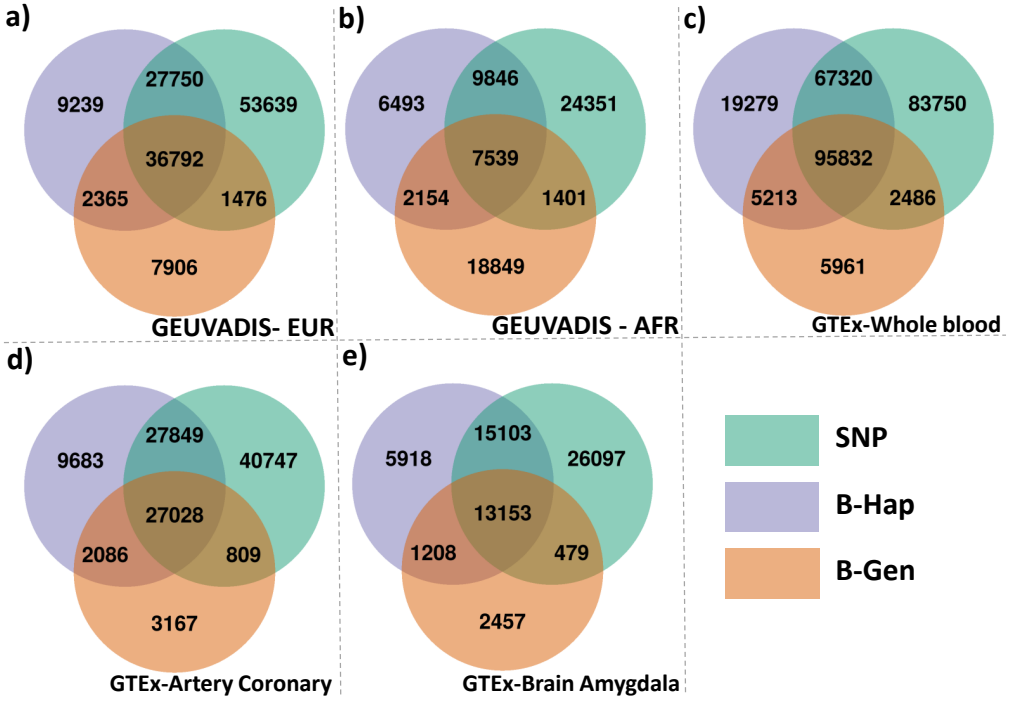



Figure 5: Venn diagram for detected significant associations from GTEx and GEUVADIS datasets. a) and b) are obtained from analysing 22,466 and 22,474 genes within 22 chromosomes of EUR (373 individuals) and AFR (89 individuals) populations from GEUVADIS dataset, respectively. c) is obtained from analysing 19,696 genes of GTEx-whole blood tissue (670 individuals). d) is obtained from analysing 23,735 genes of GTEx-artery coronary tissue (213 individuals). c) is obtained from analysing 23,268 genes of GTEx-brain amygdala tissue (129 individuals). These associations are considered significant as their BH-corrected p-value $< 0.05$ and permutation-based p-value $< 0.015$.

eQTL results those of previous meta-analyses [28, 29] finding our 11 eQTLs overlap with significant findings reported in at least one of these substantially larger studies, highlighting the increased power of our proposed approach for certain eQTLs.

We also searched for eQTLs that were detectable using haplotype but not single SNPs that could be replicated across tissues in the GTEx dataset. In total (including haplotypes that are detected using single SNPs), we find 13,573 common associations (1,312 unique genes) across whole blood (187,644 associations), artery coronary (66,646 associations), and brain amygdala (35,382 associations) tissues. The averaged p-value for these associations respectively to the mentioned tissue order is $1\times10^{-4}$, $4\times10^{-4}$, and $7\times10^{-4}$ indicating that many of these are highly significant. Of these, 729, 1085, and 947 eGenes were only detected by haplotype-based eQTL applied to whole blood, artery coronary, and brain amygdala, respectively. 11 eGenes of them replicate in across at

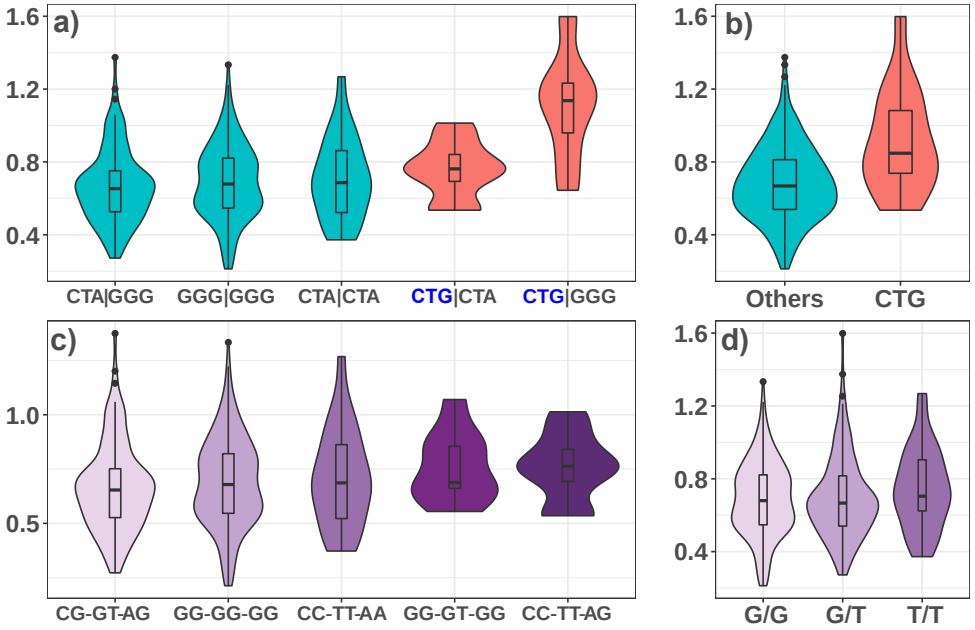



Figure 6: Comparison of gene expression distribution of *USP46-AS1* for a haplotype block using different representations. The block consists of three intronic SNPs (rs7657404 (MAF=0.34), rs7698053 (0.3) and rs7688816 (0.33)) on chr4:53186299-53186401. The different representations are a) phased haplotypes (q-value=0.017). b) a simplification of the strongest effect haplotype (CTG, frequency 2.2%) vs all others. c) all genotypes (q-value =0.99). d) the most significant SNP within the block (rs7657404, q-value = 0.9). p/q-values are calculated after eliminating haplotypes/genotypes with frequency $\leq 0.02$ but all trends remain the same if these are included.

least one other tissue type.

These findings demonstrate that our haplotype approach is able to uncover novel eQTLs that are undetectable by single SNP approaches, with a subset replicating across either dataset or tissue. These results further highlight the utility of phase-aware haplotypes for eQTL analysis.

# 4 Conclusion

In this study, we propose *eQTLHap*, a haplotype-based eQTL approach at a block scale that serves as a complementary analysis to genotype-based eQTL.

Our results show that haplotype-based eQTL outperformed other approaches for eQTL when the causal genetic architecture comprises multiple SNPs. According to the results obtained in this study, the three approaches of eQTL (based on SNP, block's genotype and block's haplotype) agreed on a large proportion of the detected associations. At the same time, each approach captured a

unique subset of the associations that have not been detected by other approaches. This observation shows that the different approaches complement each other and can provide a comprehensive eQTL scan when applied together.

To the best of our knowledge, this is the first study that investigates how phasing errors affect the results of downstream eQTL analysis. Experiments applied to synthetic haplotype and gene expression datasets demonstrated the small impact of phasing errors on TPR of downstream eQTL findings (SE $< 2.5\%$). However, there was variation in all reported significant associations demonstrate by increased false positive and negative rates when SE increased from 0% to 2.5%. This impact can be mitigated by improving phasing accuracy using consensus estimators [12] or phasing the data multiple times, applying eQTL analysis to each and keeping the stable findings.

The higher TPR obtained by haplotype-based eQTL can be justified by the less conservative MTC applied to block assessment compared to single SNPs as block count is substantially less than SNP count. However, the fact that haplotype-based eQTL outperformed block's genotype eQTL analysis in most of the experiments demonstrates that this enhanced performance is also associated with including haplotype information in the analysis, as both assessments are applied to the same blocks.

The findings when applying eQTLHap to real dataset demonstrate the efficacy of this approach as there was a large agreement with standard SNP-based eQTL approaches. In addition, several results were only revealed when considering haplotype information which replicated in both GTEx and GEUVADIS datasets as well as independent analyses available through previous meta-analyses. An interesting avenue of future research would be to explore the properties of these haplotype-based eQTLs to understand whether they represent examples of phase-specific regulation of expression or whether the findings are due to changes in the statistical test. In either case, the findings in this study highlight the potential of haplotype eQTL analysis to uncover eQTLs that have been missed through standard SNP based analysis.

eQTLHap provided in this study is simplified to accept several configurations in order to control and customise the analysis based on the user's preference. eQTLHap allows applying SNP-based, block's genotype/haplotype-based analysis separately and combined (with/without covariates). It accepts any kind of haplotype blocks. MTC can be applied using several standard methods, as well as permutation-based correction. Finally, all thresholds and options used for filtration and other purposes can be tuned.

## Acknowledgements

The authors gratefully acknowledge the GEUVADIS study. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data

used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v8.p2 on 12/13/2018.

## Funding

This work was supported by MRS scholarship [103500], the University of Melbourne and a top-up scholarship, Data61 awarded to ZB.

## References

[1] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[2] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[3] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506, 2013.

[4] Andrew E Jaffe, Richard E Straub, Joo Heon Shin, Ran Tao, Yuan Gao, Leonardo Collado-Torres, Tony Kam-Thong, Hualin S Xi, Jie Quan, Qiang Chen, et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci*, 21(8):1117–1125, 2018.

[5] Alexandra C Nica and Emmanouil T Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362, 2013.

[6] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[7] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[8] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, et al. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249, 2014.

[9] Robert Brown, Gleb Kichaev, Nicholas Mancuso, James Boocock, and Bogdan Pasaniuc. Enhanced methods to detect haplotypic effects on gene expression. *Bioinformatics*, 33(15):2307–2313, 2017.

[10] Sophie Garnier, Vinh Truong, Jessy Brocheton, Tanja Zeller, Maxime Rovital, Philipp S Wild, Andreas Ziegler, Thomas Munzel, Laurence Tiret, Stefan Blankenberg, et al. Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. *PLoS genetics*, 9(1):e1003240, 2013.

[11] Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J Topol, and Nicholas J Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215, 2011.

[12] Ziad Al Bkhetan, Justin Zobel, Adam Kowalczyk, Karin Verspoor, and Benjamin Goudey. Exploring effective approaches for haplotype block phasing. *BMC bioinformatics*, 20(1):540, 2019.

[13] Dingge Ying, Mulin Jun Li, Pak Chung Sham, and Miaoxin Li. A powerful approach reveals numerous expression quantitative trait haplotypes in multiple tissues. *Bioinformatics*, 34(18):3145–3150, 2018.

[14] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[15] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):69, 2017.

[16] Stacey B Gabriel, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002.

[17] Ning Wang, Joshua M Akey, Kun Zhang, Ranajit Chakraborty, and Li Jin. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *The American Journal of Human Genetics*, 71(5):1227–1234, 2002.

[18] Jeffrey C Barrett, B Fry, JDMJ Maller, and Mark J Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2004.

[19] Sun Ah Kim, Chang-Sung Cho, Suh-Ryung Kim, Shelley B Bull, and Yun Joo Yoo. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated snps. *Bioinformatics*, 34(3):388–397, 2018.

[20] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[21] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5), 2016.

[22] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[23] Ziad Al Bkhetan, Gursharan Chana, Kotagiri Ramamohanarao, Karin Verspoor, and Benjamin Goudey. Evaluation of consensus strategies for haplotype phasing. *bioRxiv*, 2020.

[24] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.

[25] Amy L Williams, Nick Patterson, Joseph Glessner, Hakon Hakonarson, and David Reich. Phasing of many thousands of genotyped samples. *The American Journal of Human Genetics*, 91(2):238–251, 2012.

[26] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500, 2012.

[27] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[28] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, et al. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature genetics*, 45(10):1238–1243, 2013.

[29] Rick Jansen, Jouke-Jan Hottenga, Michel G Nivard, Abdel Abdellaoui, Bram Laport, Eco J de Geus, Fred A Wright, Brenda WJH Penninx, and Dorret I Boomsma. Conditional eqtl analysis reveals allelic heterogeneity of gene expression. *Human molecular genetics*, 26(8):1444–1451, 2017.

## 5.7 Supplementary methods

### 5.7.1 Haplotype block determination and representation

Haplotype blocks are determined based on linkage disequilibrium (LD). PLINK software (Purcell et al., 2007) that implements Confidence Interval "CI" algorithm (Gabriel et al., 2002) is utilised for this purposes using its default parameters that limit the calculation to SNP pairs within 20kbp distance. A pair is considered in a strong LD if the bottom of the 90% D-prime confidence interval is greater than 0.70, and the top of the confidence interval is at least 0.98. These configurations can be changed using PLINK options listed at this link `https://www.cog-genomics.org/plink/1.9/ld#blocks`. All SNPs between the first and last SNPs of each block are considered as long as their MAF is greater than 0.01. We used this approach as it is reported to maintain high accuracy of the phased haplotypes at block scale (demonstrated in the third chapter).

Previous studies used specific SNPs to represent the haplotypes for eQTL and GWAS studies termed haplotype tagging SNPs (htSNPs) (Garnier et al., 2013; Trégouët et al., 2009). In this study, different haplotype templates based on different htSNPs were determined within each block through an iterative procedure described below:

1. SNPs with $MAF = 0$ are removed as they do not add to the associations.
2. Pearson correlation is calculated for each pair of SNPs within the block using phased alleles.
3. For each pair of SNPs with correlation equal to 1, an arbitrary SNP is eliminated as such SNPs do not affect haplotype diversity within the block.
4. In an iterative procedure that stops when the unique haplotypes based on the chosen htSNPs are $\geq \alpha$ of the unique haplotypes based on all SNPs, do the following:
   (a) Pick the pair of SNPs with the highest correlation.
   (b) Remove one SNP of this pair that has the highest correlation with the remaining SNPs. As we want to keep the most informative SNPs that differentiate between haplotypes.

Four different thresholds were used on simulated data: 70% and 80%, 90% and 100% (complete haplotype).

### 5.7.2 Gene expression simulation

Gene expression data were simulated for comprehensive evaluation and comparison between different approaches for eQTL analysis similar to the studies (Brown et al., 2017). 11 different scenarios were used for the simulation using the following general equation:

$$\begin{aligned} y &= \beta + \epsilon \\ where \quad \epsilon &\sim \mathcal{N}(0, var(\beta)\frac{1-\sigma^2}{\sigma^2}) \end{aligned} \tag{5.1}$$

$\sigma^2$ represent the proportion of the variance in the gene expression that can be explained by a specific causal genetic architecture(Brown et al., 2017). In this study, 10 different values were investigated (from 1% to 10%, by 1% step). Results for $\sigma = 5\%$ were reported in the manuscript. $\beta$ represents the causal architecture encoding that can be calculated for each simulation type as follow:

1. **Null:** This simulation represent the cases where there is not genetic architecture for influencing the gene expression.

$$\beta = 0 \quad and \quad \epsilon \sim \mathcal{N}(0, 1) \tag{5.2}$$

   With respect two this causal architecture we simulated two scenarios: Common causal SNPs when the SNPs are chosen when their MAF $\geq 0.1$ and for rare causal SNPs with $0.02 \leq$ MAF $\leq 0.05$. common and rare SNPs used in all simulations are determined using the same thresholds mentioned here.

2. **Single causal SNP:** the encoding of the causal SNP used for the simulation is the scaled minor allele dosage of the SNP (0, 1, and 2).

$$\beta = scale(g_i) \quad ; \quad g_i \in [0, 1, 2] \tag{5.3}$$

   With respect two this causal architecture we simulated two scenarios: Common causal SNPs when the SNPs are chosen when their MAF $\geq 0.1$ and for rare causal

SNPs with $0.02 \leq$ MAF $\leq 0.05$. common and rare SNPs used in all simulations are determined using the same thresholds mentioned here.

3. **Additive impact of two SNPs:** The encoding of causal architecture is calculated based on the genotypes of a random pair as follow:

$$\beta = scale(g_i) + scale(g_k) \quad ; \quad g_i \quad and \quad g_k \in [0, 1, 2] \tag{5.4}$$

Three scenarios were simulated with respect to this architecture: two common SNPs, two rare SNPs, and a pair of common and rare SNPs.

4. **SNPs interactions:** The encoding of causal architecture is calculated based on the genotypes of a random pair as follow:

$$\beta = scale(g_i) * scale(g_k) \quad ; \quad g_i \quad and \quad g_k \in [0, 1, 2] \tag{5.5}$$

Three scenarios were simulated with respect to this architecture: two common SNPs, two rare SNPs, and a pair of common and rare SNPs.

5. **Causal haplotype:** the encoding of the causal haplotype is calculated similarly to the single SNP architecture. Instead of the minor allele dosage, we used the dosage of a specific haplotype within the paternal and maternal copies of each block. Blocks are determined using plink software as mentioned in 5.7.1 Haplotype block determination and representation. Haplotypes are chosen not to be highly correlated with any single SNP within the same block ($r < 0.8$).

$$\beta = scale(h_i) \quad ; \quad h_i \in [0, 1, 2] \tag{5.6}$$

Two scenarios were simulated for this architecture: rare and common haplotypes using the same thresholds for rare and common SNPs.

When the causal architecture of any simulation involves two SNPs ($S1$, $S2$), either additive impact or interaction, the following conditions are satisfied:

1. Genomic distance between the SNPs is $\leq 7.5$ kbp.
2. Squared Pearson correlation between $S1$ and $S2$ is $< 0.8$.

3. With respect the encoding of the pair $\beta$, Pearson correlation between $S1$ and $\beta$ as well as $S2$ and $\beta$ is $< 0.8$.

Pair simulations were also repeated with the same conditions above, but forcing the pairs to be within the same haplotype block. Haplotypes for simulations were also picked when the Squared Pearson correlation between the causal haplotype and all SNPs within the same block is $< 0.8$. These constraints avoid the scenarios where these simulations are equivalent to single SNP simulations.

### 5.7.3 Comprehensive eQTL analysis

eQTL analysis was conducted in this study for each gene-block pairs with respect to the block's genotype encoding, haplotype encoding and each SNP separately as follow:

1. **Gene expression-SNP association** The association was assessed between the gene expression and each SNP with a block separately. Each SNP is assumed to have an additive impact encoded based on the minor allele dosage (0, 1, and 2) (illustrated in Figure 5.1 b)). Only SNPs with MAF $> 0.01$ are considered for this assessment. A linear regression model (Equation 5.7) is used to represent the relation. With respect to the example mentioned in Figure 5.1 b) there will be four different models as follows: $y \sim s_1$, $y \sim s_2$, $y \sim s_3$, $and\, y \sim s_4$. The significance of each relation is calculated using t statistics and p-value.

$$y = \beta s + \alpha \tag{5.7}$$

   Where $s$ is the dosage of the minor allele within the SNP , $s \in (0, 1, 2)$ .

2. **Gene expression-block's genotype association** The genotypes of SNPs are combined together into one value representing the genotype of the whole block as illustrated in Figure 5.1 c). Analysis of variance (ANOVA) model is used to represent the relation between gene expression and genotype-based encoding of a block. F-test and p-values are calculated based on the ANOVA model and reported for the association. The ANOVA model is equivalent to multiple linear regression model:

$$y = \sum_{i=1}^{m} \beta_i g_i + \alpha \tag{5.8}$$

Where $m$ is the number of unique genotypes within the block. $g_i \in (0, 1)$ represents whether the individual carries the genotype $g_i$ or not. Rare genotypes within each block were eliminated from the model (genotype frequency $\leq 0.02$). Considering the same example in Figure 5.1 c), the model is

$$y = \beta_1 g_1 + \beta_2 g_2 + \beta_3 g_3 + \beta_4 g_4 + \alpha \tag{5.9}$$

For a block's genotype $g$, $g_1 = I(g = 1210)$, $g_2 = I(g = 1220)$, $g_3 = I(g = 1211)$ and $g_4 = I(g = 2200)$.

3. **Gene expression-block's haplotype association** Haplotypes are encoding similarly to bag-of-words representation using in text mining. unique haplotypes within a block and across all individuals are determined then individual's block is represented using the dosage of each haplotype within both homologous chromosome copies as illustrated in Figure 5.1 d). A multiple linear regression model is used to assess the gene expression as a response to the additive model of the haplotypes within each block.

$$y = \sum_{i=1}^{m} \beta_i h_i + \alpha \tag{5.10}$$

Where $m$ is the number of unique haplotypes within the block. $h_i$ is the dosage of the haplotype $i$, where $h \in (0, 1, 2)$. F statistic and p-value is calculated for the whole model. Rare haplotpes within each block were eliminated from the model (haplotype frequency $\leq 0.02$). Considering the example in Figure 5.1 d), the model is

$$y = \beta_1 h_1 + \beta_2 h_2 + \beta_3 h_3 + \alpha \tag{5.11}$$

For block's hapotypes $\bar{h_1}$ and $\bar{h_2}$:
$h_1 = I(\bar{h_1} = 0110) + I(\bar{h_2} = 0110)$,
$h_2 = I(\bar{h_1} = 1100) + I(\bar{h_2} = 1100)$, and
$h_3 = I(\bar{h_1} = 0111) + I(\bar{h_2} = 0111)$.

Before conducting any statistical assessment, SNPs dosage, haplotype dosage, and gene expression were standardised as used in a well-known eQTL analysis tool (Matrix eQTL (Shabalin, 2012)).

### 5.7.4 Consideration of covariates

It is popular to account for covariates in eQTL analysis such as individual's gender, imputed SNPs, principle components determined for individuals genotypes (to avoid population stratification's as explained in background 2.2 Quality control procedure and data preparation). For such cases, the regression models above become:

$$y = \beta s + \sum_{j=1}^{c} \gamma_j v_j + \alpha \tag{5.12}$$

$$y = \sum_{i=1}^{m} \beta_i g_i + \sum_{j=1}^{c} \gamma_j v_j + \alpha \tag{5.13}$$

$$y = \sum_{i=1}^{m} \beta_i h_i + \sum_{j=1}^{c} \gamma_j v_j + \alpha \tag{5.14}$$

where $c$ is the covariates number. The corrected degrees of freedom for the model decreases by $c$ to account for the added covariates.

### 5.7.5 Technical implementation of the statistical assessment

Statistical assessment for eQTL analysis is time consuming due to a large number of possible gene expression-genomic locus pairs. For fast execution, we adopted Matrix eQTL approach (Shabalin, 2012) that relies on matrix operations.

For a haplotype block for $n$ individuals and covariate matrix ($v$) of $c$ covariates, the applied algorithm (adapted from Matrix eQTL to suit blocks instead of SNPs) is described below:

1. Create bag of haplotype (*boh*) data structure for the block as a list of $m$ one-row matrices of a length $n$.
2. Fill each matrix with the dosage of its associated haplotype within the paternal and maternal copies.
3. Center all matrices: gene expression ($y$), each matrix of the bag of haplotype in (1), and the covariates ($v$) to remove the intercept ($\alpha$) from the model.

4. Orthogonalise the gene expression with respect to the covariates.

$$\tilde{y} = y - \langle y,\, v\rangle v \tag{5.15}$$

   Where $\langle y,\, cov\rangle$ is the cross product of $y$ and $cov$.

5. Orthogonalise each matrix ($h_i$) of *hob* with respect to the covariates and other matrices.

$$\ddot{h_i} = h_i - \langle h_i,\, v\rangle v \quad ;\, i = 1, 2, \ldots, m. \tag{5.16}$$

$$\tilde{h_i} = \ddot{h_i} - \langle \ddot{h_i},\, \tilde{h_j}\rangle h_j \quad ;\, i = 1, 2, \ldots, m; \quad j = 1, \ldots, i-1. \tag{5.17}$$

6. Standardise each matrix ($\tilde{h_i}$) of *hob*.

7. Calculate test statistic $R^2$:

$$R^2 = \sum_{i=1}^{m} \langle \tilde{y},\, \tilde{h_i}\rangle^2 \tag{5.18}$$

8. Calculate F-test then pvalue based on $R^2$

$$\text{F-test} = \frac{(n - m - c)R^2}{m(1 - R^2)} \tag{5.19}$$

In case no covariates provided, the same algorithm is applied by ignoring the steps that involve the covariates. For SNP-based eQTL, the same algorithm as Matrix eQTL is applied considering the simple linear regression model.

The same algorithm can be applied to assess the association between gene expression and the genotype of the block. The only difference is to change boh matrices to include the dosage of the genotypes.

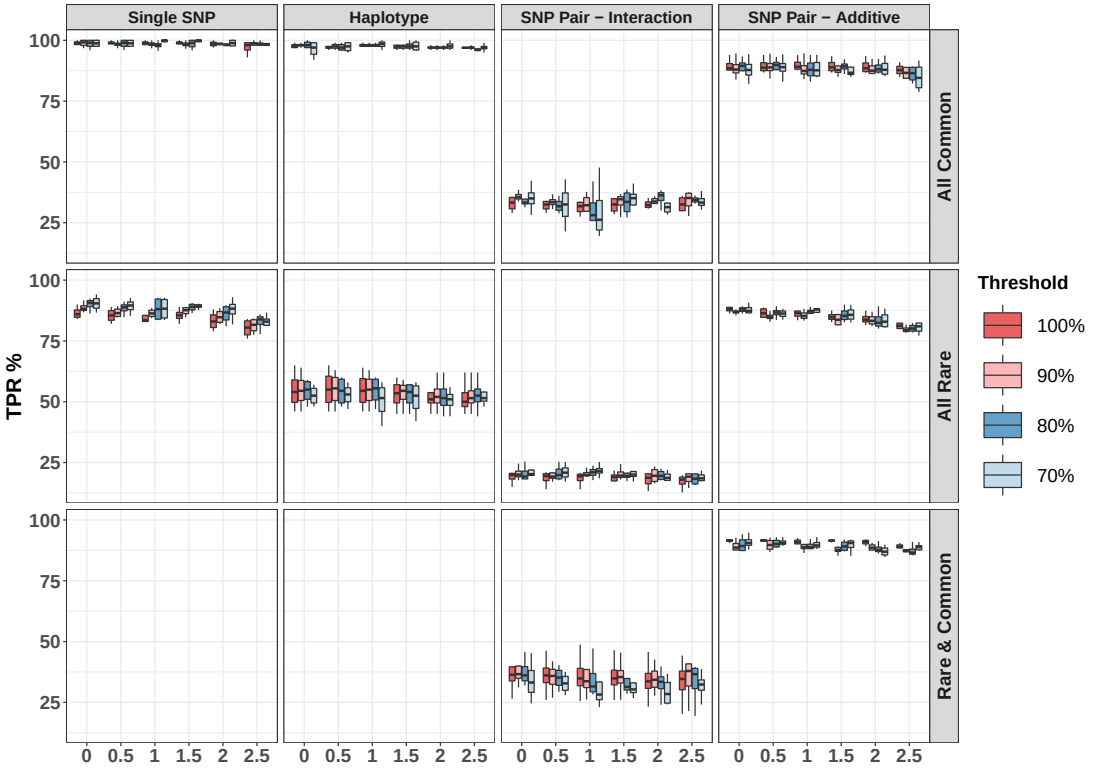
## 5.8 Supplementary results

### 5.8.1 Different templates for haplotype blocks

We investigated different haplotype templates for association assessment to find the template that leads to the best results. For this purpose, htSNPs were chosen in a way they represent 70 %, 80% and 90% of all haplotypes determined using all SNPs

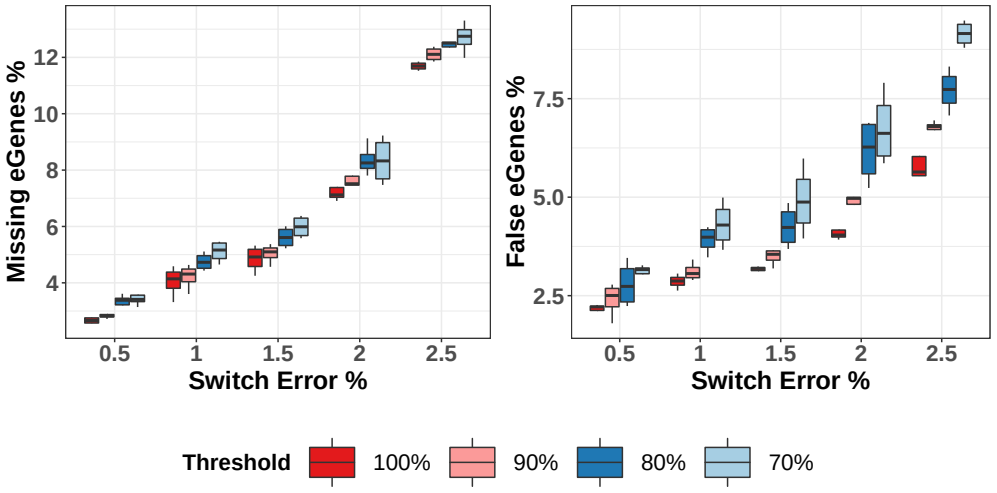
within each block. Using simulated data with different switch error rates, we found that different templates obtained similar TPR as illustrated in figure 5.7. For further

**Figure 5.7: TPR of haplotype-based eQTL using different haplotype templates**. TPR calculated haplotype templates representing 100%, 90%, 80% and 70% of blocks haplotypes with respect to different causal architecture and causal frequency. Each plot represents a TPR comparison for different switch error rate from 0 to 2.5%. Results are summarised for all causal architectures mentioned in section 2.5 Gene expression simulation.




investigation, we compared results' variation for different SEs and with respect to each template. Compared to the associations detected on haplotypes without errors (SE = 0%), using all SNPs within a block maintained more similar results than other templates as illustrated in Figure 5.8. Both missing eGenes (only reported on data with SE = 0%) and False eGenes (only reported on data with errors) increase when the SNPs within each block decrease. Moreover, The difference between these templates increases with the switch error rate. These observations encouraged to use 100% template for the majority of the results in this chapter.

**Figure 5.8: Missing and False eGene for different haplotype templates.** A comparison of the percentage of Missing eGenes and false eGenes for different haplotype templates and switch errors. These results represent a comparison to the eGenes obtained on the data without errors (SE = 0%). Results are summarised for all causal architectures mentioned in section 2.5 Gene expression simulation.




### 5.8.2 Performance of eQTL approaches with different gene expression heritablity

In this section, we report the performance of eQTL approaches with respect to different values of $\sigma$ from 0 to 0.1 where $\sigma$ is the proportion of the variance in the gene expression that can be explained by a specific causal genetic architecture. This experiment was applied to genotype data obtained from the 1000 genome project for 373 individuals from European population and gene expression data simulated for the genes ENSG00000127074, ENSG00000162441, ENSG00000196539, ENSG00000218510, ENSG00000198468 genes as used in the section 2.6 Simulations to compare different eQTL approaches. As expected the TPR increases with $\sigma$ as illustrated in Figures 5.9, 5.10 and 5.11 that represent causal architectures with different frequencies.

eQTL based on block's genotype obtained the minimal TPR for all experiments with an exception the one simulated based on haplotypes. SNP-based eQTL outperformed all other approaches when the causal architecture is a single SNP regardless of the $\sigma$. Haplotype-based eQTL analysis is superior to other approaches when the causal architecture involves multiple SNPs. When the causal is a pair of common SNPs, the improvement in TPR obtained by haplotype-based eQTL increases with $\sigma$ as illustrated

in Figures 5.9, 5.10 and 5.11. TPR patterns for all approaches are similar regardless of the frequency of causal architecture.

**Figure 5.9: TPR of eQTL approaches for different common causal architectures.**

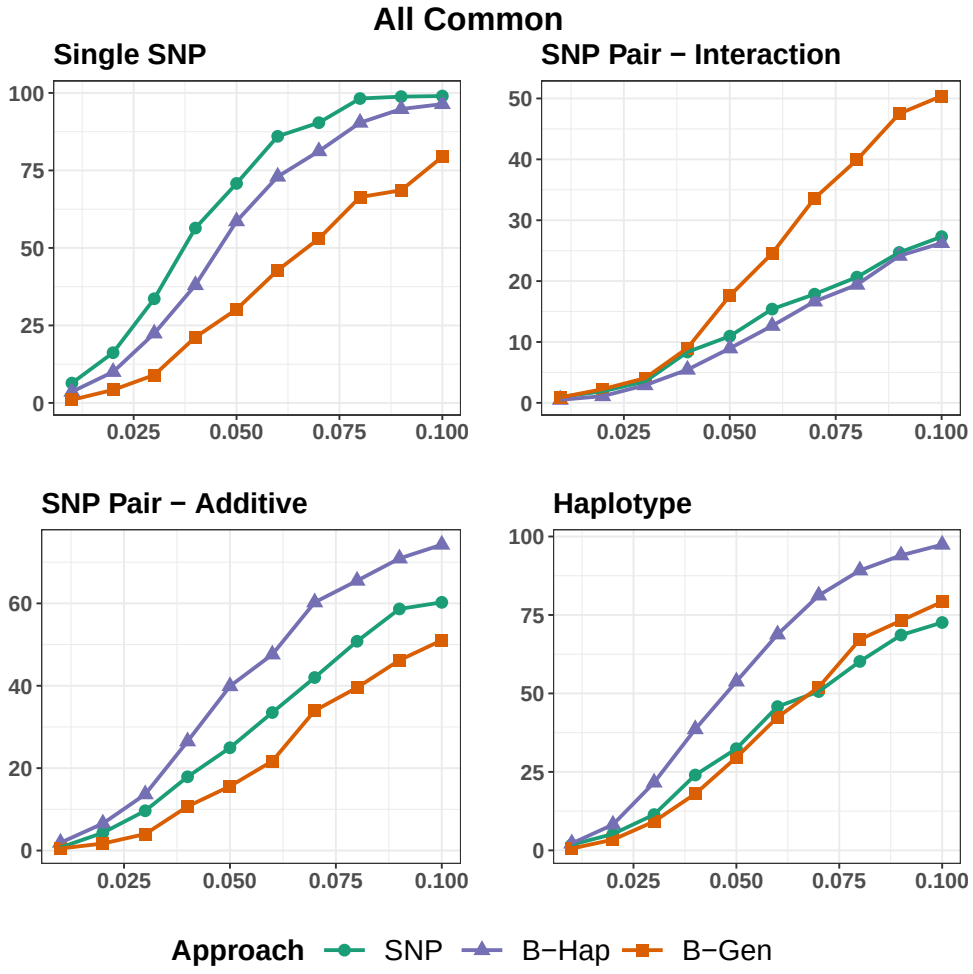


**Figure 5.10: TPR of eQTL approaches for different rare causal architectures.**

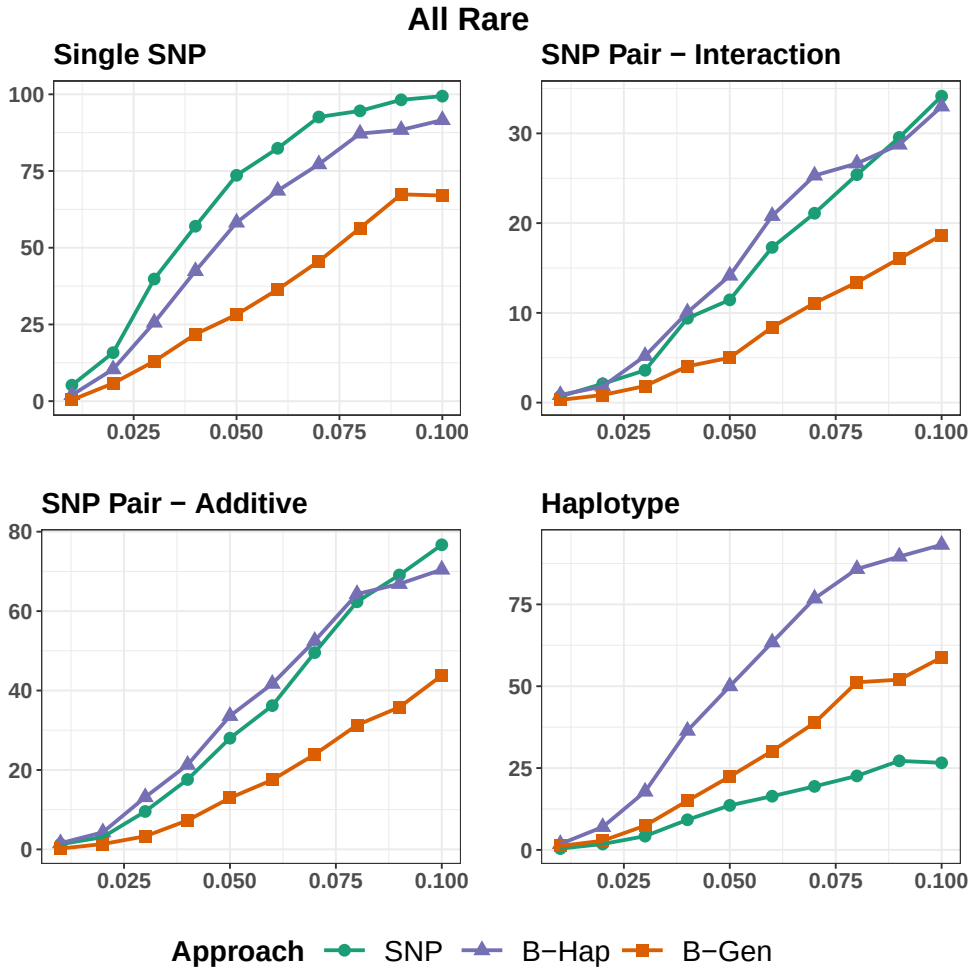
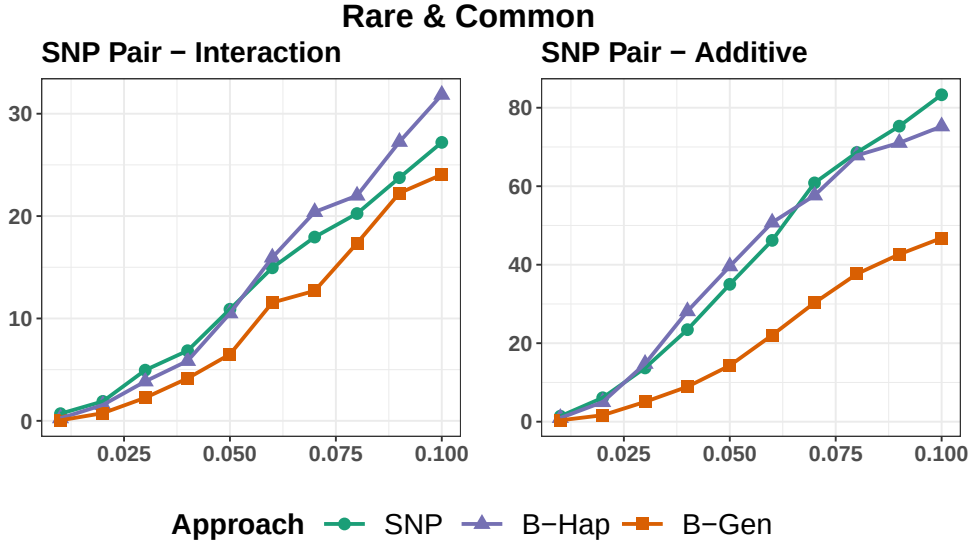
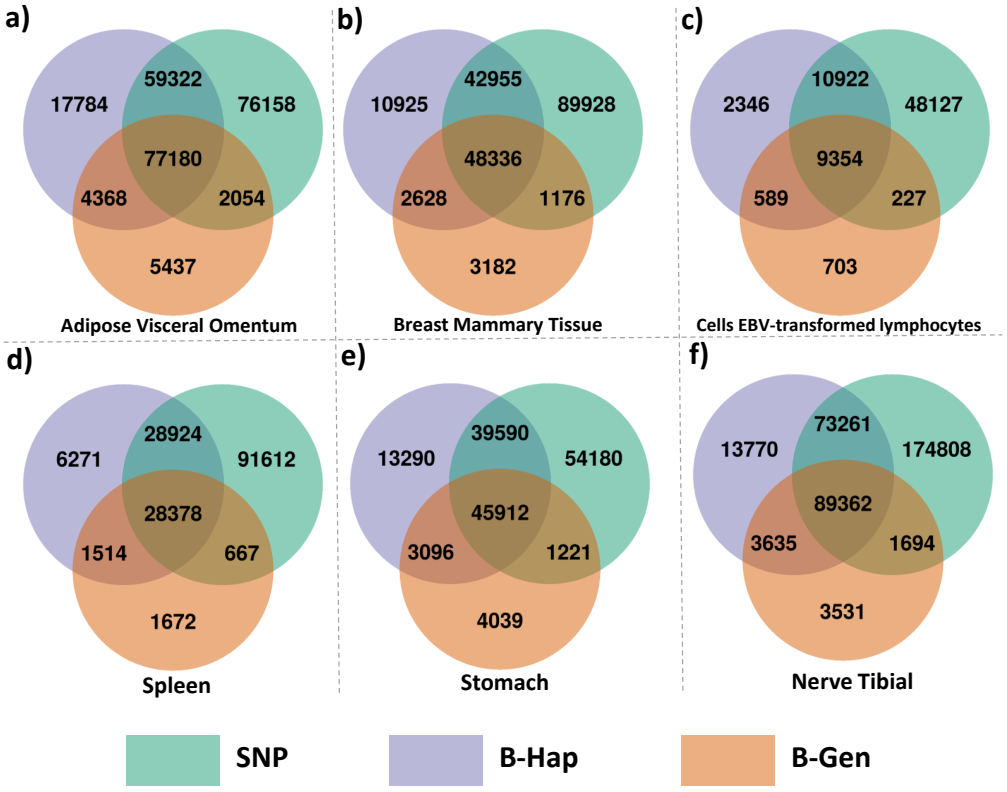



**Figure 5.11: TPR of eQTL approaches for causal architectures based on a pair of rare and common SNPs.**



### 5.8.3 The application of eQTLHap to other tissues from GTEx dataset

**Figure 5.12: Venn diagram for detected significant associations from different tissues provided by GTEx datasets.** a) is obtained from analysing 24,724 genes of GTEx-adipose visceral omentum tissue (469 individuals). b) is obtained from analysing 25,849 genes of GTEx-breast mammary tissue (396 individuals). c) is obtained from analysing 22,759 genes of GTEx-cells EBV-transformed lymphocytes tissue (147 individuals). d) is obtained from analysing 25,479 genes of GTEx-spleen tissue (227 individuals). e) is obtained from analysing 24,290 genes of GTEx-stomach tissue (324 individuals). f) is obtained from analysing 25,873 genes of GTEx-nerve tibial tissue (532 individuals).



# Chapter 6

# Discussion

Accurate and comprehensive investigation of disease-genomic variation associations requires the interrogation of different representations of the genome. Such a comprehensive analysis is essential when the genetic basis underlying specific disease is unknown. Therefore, there have been numerous genetic investigations conducted on different datasets for several diseases considering individual SNPs (Manolio, 2017), epistasis (Niel et al., 2015), SNP combination in blocks (Howard et al., 2017; Trégouët et al., 2009). In most of these studies, the genotypic representation of the genome was prominently used in these studies. Such representation accounts for the different alleles within a SNP yet it ignores their location on the homologous copies of chromosomes. Fewer studies considered the haplotypic representation of the genome in the analysis, that is, investigating a pair of allele sequences within the homologous chromosome copies. Utilising haplotype information for genetic investigations has been reported to empower traditional methods used for several genetic problems such as Expression quantitative trait loci (eQTL) (Brown et al., 2017; Garnier et al., 2013) and genome-wide association studies (GWAS) (Browning, 2008).

In most of haplotype-based analyses, haplotype information is derived through computational methods called haplotype phasing or estimation methods (Browning and Browning, 2011). Therefore, researchers focused on developing accurate and efficient phasing tools in the last two decades, especially with the limited experimental methods that could not provide this information before. Despite the recent developments in genome sequencing methods (especially next and third-generation sequencing) that are now capable of obtaining haplotypes of short and long reads, the significance of

computational haplotype phasing can not be negated as there are a substantial number of different genotype datasets that have been published before and require phasing for different purposes such as genotype imputation, GWAS and eQTL.

One of the most popular approaches to use haplotype information in studies concerning genetic association with phenotypes is to computationally phase the genotypes into haplotypes then partition the estimated haplotypes into blocks. After that, statistical tests are conducted to assess the relationship between the blocks and the phenotype which can be categorical as in case/control studies or continues as in eQTL studies (as described in chapter 3 and 4). Performing precise haplotype-based analysis requires paying great attention to phasing method as well as the partitioning approach used as both factors have an impact on the accuracy of phased haplotypes within the determined blocks and thereby they significantly influence the findings and results of the haplotype-based analyses.

The focus in this PhD study is on improving the accuracy of computational haplotype phasing and exploring optimal ways to utilise the estimated haplotypes for precise haplotype-based investigations. While we use genotype imputation and haplotype-based eQTL analysis as two case studies, we believe that approaches used and findings reported have the potential to lead to more accurate other downstream applications such as GWAS.

We demonstrated in this PhD that accurate and stable haplotype phasing can be achieved by combining multiple phasing tools or multiple iterations of one non-deterministic tool within a consensus estimator. The decision on the optimal structure to use is related to the characteristics of the data being processed such as sample size and SNP density. Both structures are more accurate than any individual tool. This improvement was assessed for both standard metrics of haplotype phasing evaluation as well as by the improvements achieved with the downstream application of phased haplotypes. In the context of genetic association analysis, we show that haplotype information significantly adds to the ability to detect novel associations especially for particular underlying genetic architectures that comprise multiple variants. We believe that haplotype-based analysis complements SNP-based analysis and that both analyses applied together provide a more comprehensive scan that increases the chance to elucidate the genetic basis

of disease aetiology. These findings confirm the important role and the additional enhancement that haplotype information can add to typical methods used in the genetic field. We also provided in this PhD tools to support other researchers interested in investigating haplotype information in genetic investigations, thus reducing the time taken for analysis.

We have structured this PhD in three main studies where we started by exploring optimal ways to determine haplotype blocks for investigations based on haplotype blocks such as eQTL and GWAS. This study encouraged constructing a consensus haplotype estimator to improve phasing accuracy. Therefore, in the second study, we investigated the efficacy of this consensus approach generally and with respect to genotype imputation. In the third study, we utilised all findings from the first two studies in order to develop a new method for haplotype-based eQTL analysis. Observations, findings, approaches and potential future directions related to these studies are discussed below.

## 6.1 Exploring effective approaches for haplotype block phasing

In this study, we demonstrated how both haplotype block determination method and the phasing approach jointly impact the error rate within haplotype blocks. We demonstrated that errors obtained by well-known phasing tools (EAGLE2, HAPI-UR, SHAPEIT2, BEAGLE, IMPUTE, MaCH, and fastPHASE) occur in different loci which encourages to construct a consensus estimator from multiple tools. However, we showed that the tools included in the consensus estimator have a significant impact on accuracy. Our results showed that a consensus estimator of SHAPET2, EAGLE2, and BEAGLE obtains consistently the best results compared to other combinations and any single tool.

Furthermore, we report a novel evaluation of phasing accuracy with respect to two different haplotype block determination methods (using sliding window and LD-base determination) and introduce a new metric for this purpose (Incorrect haplotype block percentage (IHBP)). We reported the impact of sliding window width on the obtained accuracy. We also explained the trad-off between the highest accuracy obtained when dealing with LD-based blocks and the comprehensive scan when dealing with sliding window.

The key messages of this study to the reader are:

- Phasing accuracy can be improved using a consensus estimator of multiple tools. However, if any constituent tool has a high error rate, the consensus approach does not necessarily outperform a single accurate tool. It is very important to decide on the constituent tools wisely.
- A consensus estimator of SHAPEIT2, EAGLE2 and BEAGLE not only reduced switch error by 14% but also increased the stability of the results.
- Determining haplotype blocks based on LD leads to low error rate within the blocks, however, it is less comprehensive than a sliding window.
- When applying a sliding window, it is very important to consider a reasonable window width, as it has a significant impact on the accuracy.

Future work related to the experiments and findings reported in this chapter:

- Evaluation to other haplotype block determination methods such as Four gamete (Wang et al., 2002) and Solid spine (Barrett et al., 2004).
- Genome-wide haplotype association analysis relies on haplotype blocks (Howard et al., 2017; Lv et al., 2017; Shang et al., 2015; Trégouët et al., 2009), therefore, this evaluation can be extended to assess the impact of the accuracy at block scale on the downstream association analysis.
- Similarly to the idea of a consensus approach for phasing, haplotype blocks can be determined through a consensus agreement. Blocks boundaries can be determined within SNPs when phasing tools disagree on phasing. This idea is credited to an anonymous reviewer.

## 6.2 Evaluation of consensus strategies for haplotype phasing

In this study, we demonstrate that improvements in phasing accuracy are observed through the combination of different phasing tools or with multiple iterations of a single

non-determining tool. We show that the consensus remains the most accurate approach when data characteristics such as sample size, SNP density, and minor allele frequency are varied, consistently outperforming any of the considered tools individually. The computational cost of the accuracy improvement is reported and discussed to show possibilities to mitigate this extra time required.

We also assess the impact of phasing accuracy on genotype imputation, one of the major applications of phasing. Up to the best of our knowledge, this is the first evaluation of the influence of phasing on downstream genotype imputation that considers all phasing and genotype imputation tools available in the widely-used Sanger and Michigan imputation servers as well as reporting phasing and imputation accuracy at individual scale and whole dataset.

The key messages of this study to the reader are:

- The highest accuracy gain is to construct a consensus of SHAPEIT2, SHAPEIT3, and EAGLE2 when dealing with large and high SNP density datasets, and a consensus of multiple iterations of SHAPEIT2 for small and low SNP density datasets (we recommend between 5 and 15).
- The improvement of the consensus approach is robust against factors reported to impact phasing accuracy significantly.
- With respect to genotype imputation, the improvement of phasing led to better genotype imputation. This results demonstrate the efficacy of this approach and evidenced a potential similar impact on other downstream analysis such as haplotype association assessment for GWAS and eQTL (Brown et al., 2017; Howard et al., 2017).
- We freely provide consHap tool to perform consensus haplotype construction and other supplementary tasks to handle different formats of phased data (.VCF, .haps/.sample, and .phgeno/.phind/.phsnp).

Future work related to the experiments and findings reported in this chapter:

- Investigating other approaches to construct a consensus haplotype estimator: this is an open and flexible area to work on. Different approaches from ensemble learning can be used for this purpose.

- The improvement obtained by the consensus estimator for phasing encourages to construct a consensus approach for genotype imputation as both genotype imputation and phasing tools are based on Li and Stephens model (Li and Stephens, 2003).

## 6.3 eQTLHap: a tool for a comprehensive eQTL scan considering haplotypic and genotypic effects

In this study, we demonstrated using synthetic data that including haplotype information in eQTL analysis lead to detect significant eGenes that are missed by standard SNP-based analysis. The application of the method on real data confirmed similar findings. We also report the impact of phasing errors on the downstream eQTL analysis. This novel analysis provides some confidence in the results obtained by this approach. The comparison results obtained by different representations of genomic variations (SNP's genotype, block's genotype and block's haplotypes) show the strengths and weakness of each approach.

The key messages of this study to the reader are:

- Phasing errors have an impact on eQTL analysis, however, strong associations seem to be robust against the errors. In addition, associations based on causal architecture involving common SNPs are the least affected by these errors.
- Each genomic representation led to a subset of unique results that are not captured bu other representations. The least effective approach was the one based on the block's genotype. haplotype-based approach outperformed other approaches when the causal architecture involves multiple SNPs, with higher improvement for rare SNPs.
- Haplotype-based approach is complementary to standard SNP-based eQTL analysis.
- we provide freely comprehensive eQTL tool that performs eQTL analysis at three scales: single SNP, block's genotype, and block's haplotype.

Future work related to the experiments and findings reported in this chapter:

- Different haplotype templates within determined blocks: We have explored different templates to determine the haplotypes within the blocks based on SNP correlation. Other studies used the Akaike information criterion (AIC) for this purpose. This aspect can be explored more to find an optimal haplotype representation within the blocks that can lead to better results.

- Exploring results of different haplotype block determination methods. The tool provided by this chapter facilitates the implementation of this idea.

- Exploring different ways to encode block's haplotypes/genotypes.

## 6.4 Future Directions

There are many aspects of this PhD that can be extended and investigated in more depth. In addition to the potential future work described specifically for each study above, we report here more general aspects for potential improvement and exploration.

- **Improvement of haplotype phasing and genotype imputation**
  Rapid and scalable haplotype phasing and genotype imputation have been achieved by recently developed tools such as SHAPEIT4 (Delaneau et al., 2019), BEAGLE5 (Browning et al., 2018) and IMPUTE5 (Rubinacci et al., 2020). The new approaches are using positional Burrows-Wheeler transform (PBWT). PBWT provides an efficient haplotype representation that leads to quick haplotype matching as well as less storage required. We have observed in this thesis that learning from different tools that consider different assumptions increases the accuracy more than what can be achieved by one tool (regardless whether it is used in a consensus estimator or by optimising its parameters). Developing tools that can consider these different assumptions during the estimation process can lead to high accuracy with a less computational cost.

- **Investigation of other haplotype block determination methods**
  This PhD focused on only two approaches to determine the boundaries of haplotype blocks for the evaluation (sliding window and block-based method), and only on LD-based blocks for the application of haplotype information for eQTL analysis. Other approaches can be utilised for the same purposes such as Four gamete

(Wang et al., 2002), Solid spine (Barrett et al., 2004), HaploBlocker (Pook et al., 2019), Big-LD (Kim et al., 2018), S-MIG++ (Taliun et al., 2015), MATILDE (Pattaro et al., 2008), and a sliding windows with different lengths, with overlapping and non-overlapping domains. In this thesis, all SNPs within determined blocks are considered, however, previous studies have considered a subset of these SNPs termed haplotype tagging SNPs (htSNP) (Garnier et al., 2013; Trégouët et al., 2009). After that, different combinations of these htSNPs were used to create multiple haplotypic models and the best model that minimises the scaled Akaike Information Criterion (AIC) is chosen for the assessment. This approach is mainly used when having discovery and replication phases as final haplotypic models were optimised to suit the data in the discovery phase, however, the findings are verified again using independent data in the replication stage to eliminate false positives related to the "over-fitted" assumptions based on the data in the discovery phase. Finally, the analysis can be done to focus only on regulatory regions such as promoters and enhancers. Haplotypes within these regions can be investigated to reveal any significant associations. Limiting the analysis to these regions can lead to having less conservative multiple test correction as it reduces the number of different tests. In the same time, it is known that most of the eQTL associations are related to variants in the regulatory regions.

This extension will provide a more comprehensive analysis of several ways to use haplotype information in genetic problems.

- **Extension to adapt with GWAS studies**

  Previuse studies have investigated haplotype associations with diseases (Howard et al., 2017; Lv et al., 2017; Shang et al., 2015; Trégouët et al., 2009). eQTLHap approach can be extended to consider association with phenotype within case/-control studies. The main difference is that with GWAS, the response variable is binary (0 or 1) compared to the continuous nature in the case of eQTL analysis. This difference requires to revisit the statistical assessment used to consider the categorical nature of the phenotype such as using logistic regression or chi-square test. The sample size of GWAS datasets is usually larger than the datasets used for eQTL analysis. This can be very helpful in the case of haplotype-based association analysis as in such analysis there is a less degree of freedom (possible different haplotypes are large within each block) compared to the genotype of a

single SNP in a standard GWAS method (only three different values 0, 1 and 2). Comparing GWAS results obtained by the three different encodings used in chapter 5 when applied to the same dataset can enrich our understanding of the genetic basis underlying the investigated disease or trait.

- **Haplotype-based gene expression estimation**

  Recent studies have constructed several statistical approaches that are able to estimate the gene expression data from the genotypes of individuals. PrediXcan (Gamazon et al., 2015) and FUSION (Gusev et al., 2016) are examples of these tools that are available online [1, 2]. These approaches have been proposed based on single SNPs (Gamazon et al., 2015; Gusev et al., 2016) and have been demonstrated to be effective. Up to the best of our knowledge, utilising haplotype information for this problem has not been explored before.

  Supported by the results demonstrating the efficacy of using haplotype information for detecting specific cases of eQTL (chapter 5), such an approach can be extended to construct regression models that are able to predict gene expression data using supervised learning. For example, using the same procedure applied in haplotype-based eQTL analysis in term of haplotype blocks determination and encoding. Machine learning will be applied to build regression models for gene expression imputation. The input data will be represented as a sequence of haplotype blocks. Several predictive models will be trained on phased genotypes and gene expression for the same individuals. One model will be constructed for every single gene and specific tissues. The popular regression models used for such problem are linear, elastic net and random forest.

  The predicted gene expression (for case/control study) can be used for differential gene expression (DGE) analysis in order to link gene expression with a specific phenotype. Such association can be traced back to the significant haplotype blocks used in the model training.

- **Follow-up analysis for the detected haplotype-based association**

  Significant associations detected by haplotype-based approach applied to eQTL or GWAS problem should be investigated further to understand the biological

[1] `https://github.com/hakyimlab/PrediXcan/tree/master/Software`
[2] `http://gusevlab.org/projects/fusion/`

reasons that can interpret these associations. For example, investigating the location of the detected haplotypes (on regulatory regions, intronic regions, or coding regions), and investigating epistatic interactions of haplotype blocks similarly to SNPs interaction.

- **An extension to investigate allele-specific expression (ASE)**

  One natural extension of this work is to consider the difference in the expression of the two homologous copies of each chromosome pair. However, this extension is more complicated as it requires the alignment of RNA sequencing reads to both homologous copies of chromosomes to determine the read counts of each gene copy. In addition, it is important to link each haplotype with the read counts determined for the same copy.

- **Tool improvements**

  Generally speaking, haplotype-based eQTL is a time-consuming task, therefore, investigations seeking to reduce performance time are warranted. Implementation of a matrix-based approach such as Matrix eQTL (Shabalin, 2012) can be a good starting point to achieve better performance. While the bottleneck of constructing a haplotype estimator is phasing performance, our implementation of a consensus estimator can be improved to reduce the performance time. Tools provided in this PhD study can be made more user friendly to enable usage by non-experts.

# Appendix A

# Supplementary materials for chapter 3

## A.1 Dataset details

**Table A.1: Simulated dataset details.** Heterozygosity% represents the percentage of heterozygouse SNPs to all SNPs in the dataset.

| Dataset | Population | Length | Sample | Heterozygosity% | Missing SNPs |
|---|---|---|---|---|---|
| SD1 | European | 227 | 200 | 31.51 | 2,210 |
| SD2 | European | 225 | 500 | 31.04 | 5,472 |
| SD3 | European | 1,091 | 3,484 | 30.27 | 185,464 |
| SD4 | African | 229 | 200 | 26.23 | 2,236 |
| SD5 | African | 226 | 500 | 27.99 | 5,511 |
| SD6 | African | 1,139 | 3,492 | 26.31 | 193,835 |
| SD7 | African American | 230 | 200 | 26.19 | 2,251 |
| SD8 | African American | 228 | 496 | 26.27 | 5,510 |
| SD9 | African American | 1,139 | 3,484 | 26.49 | 193,437 |
| SD10 | Asian | 201 | 500 | 34.43 | 4,893 |
| SD11 | Asian | 204 | 200 | 34.79 | 1,985 |
| SD12 | Asian | 1,075 | 3,490 | 29.31 | 182,980 |

**Table A.2: Males' chromosomes X dataset details.** Heterozygosity% represents the percentage of heterozygouse SNPs to all SNPs in the dataset.

| Dataset | Length | Sample | Heterozygosity% | Missing SNPs | Unresolved SNPs |
|---|---|---|---|---|---|
| XD1 | 199 | 200 | 33.11 | 0 | 11 |
| XD2 | 198 | 500 | 35.75 | 0 | 41 |
| XD3 | 1397 | 3500 | 36.57 | 217 | 2519 |
| XD4 | 11398 | 9992 | 35.54 | 2349 | 53288 |

**Table A.3: Real dataset details.** Heterozygosity% represents the percentage of heterozygouse SNPs to all SNPs in the dataset. First three rows represent the details of HAPMAP III individuals (children with known paternal genoimc data) , while the rest represent the details of the datasets with all individuals (Coeliac disease dataset and HAPMAP III).

| **Dataset** | **Length** | **Sample** | **Heterozygosity %** | **Missing SNPs** | **Unresolved heterozy-gouse SNPs** | **Unresolved missing SNPs** |
|---|---|---|---|---|---|---|
| Chr1 | 36,923 | 39 | 35.2 | 3032 | 98,852 | 1,846 |
| Chr6 | 31,727 | 39 | 35.14 | 3005 | 85,528 | 1,870 |
| Chr17 | 12,807 | 39 | 35.42 | 958 | 34,777 | 594 |
| Chr1 | 36,923 | 12,008 | 34.75 | 182,806 | - | - |
| Chr6 | 31,727 | 12,008 | 34.67 | 165,529 | - | - |
| Chr17 | 12,807 | 12,008 | 35.21 | 60,519 | - | - |

# Appendix B

# General Information

## B.1 Tools provided by this PhD

1. **consHap tool**: https://github.com/ziadbkh/consHap.
2. **eQTLHap tool**: https://github.com/ziadbkh/eQTLHap.

## B.2 Phasing tool and genetic maps links

1. **PHASE**: http://stephenslab.uchicago.edu/software.html#phase.
2. **fastPHASE**: http://stephenslab.uchicago.edu/software.html#fastphase.
3. **BEAGLE**: http://faculty.washington.edu/browning/beagle/beagle.html.
4. **IMPUTE**: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html.
5. **MaCH**: http://csg.sph.umich.edu/abecasis/mach/tour/input_files.html
6. **SHAPEIT2**: https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html.
7. **SHAPEIT3**: https://jmarchini.org/shapeit3/.
8. **HAPI-UR**: https://code.google.com/archive/p/hapi-ur/.
9. **EAGLE2**: https://data.broadinstitute.org/alkesgroup/Eagle/.

10. **Genetic maps 1**: https://data.broadinstitute.org/alkesgroup/Eagle/downloads.
11. **Genetic maps 2**: http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps.

## B.3 Genotype imputation tool ans servers links

1. **Beagle5**: https://faculty.washington.edu/browning/beagle/beagle.html.
2. **Minimac3**: https://genome.sph.umich.edu/wiki/Minimac3.
3. **PBWT**: https://github.com/richarddurbin/pbwt.
4. **Sanger imputation service**: https://imputation.sanger.ac.uk/.
5. **Michigan Imputation Server**: https://imputationserver.sph.umich.edu/index.html.

## B.4 Datasets links

1. **The Haplotype Reference Consortium (HRC):** Available through European Genome-phenome Archive (dataset reference: EGAD00001002729) at the link: https://www.ebi.ac.uk/ega/datasets/EGAD00001002729. The official website is: http://www.haplotype-reference-consortium.org.
2. **HapMap III**: Available publicly at: ftp://ftp.ncbi.nlm.nih.gov/HapMap/genotypes/HapMap3_r3.
3. **Coeliac disease dataset**: It is provided by the Wellcome Trust Case-Control Consortium (EGA accession: EGAS00000000057) at this link https://www.ebi.ac.uk/ega/studies/EGAS00000000057.
4. **Geuvadis dataset**: Available publicly at: https://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/.
5. **GTeX dataset**: it is provided by dbGaP at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2. The official website is: https://www.gtexportal.org/.

## B.5 Supplementary tools

1. **PLINK**: `https://www.cog-genomics.org/plink2`.

## B.6 Phasing tool running

- **BEAGLE:**

```
$ java -Xss5m -Xmx8g -jar beagle.08Jun17.d8b.jar
gt=beg/input_genotype_file.vcf.gz
map=genetic_map_file out=output_phased_haplotypes_file
```

- **EAGLE2:**

```
$ eagle --bfile=input_genotype_file --geneticMapFile genetic_map_file
--outPrefix output_phased_haplotypes_file
```

- **fastPHASE:**

```
$ plink -bfile input_genotype_file --recode fastphase
--out input_genotype_file.inp
$ fastPHASE -T20 -ooutput_phased_haplotypes_file input_genotype_file.inp
```

- **HAPI-UR:**

```
$ cp input_genotype_file original_snps.bim
$ insert-map.pl original_snps.bim genetic_map_file > input_genotype_file.bim
$ hapi-ur -p input_genotype_file -w 73 -o output_phased_haplotypes_file
```

- **IMPUTE:**

```
$ plink -bfile input_genotype_file
--recode --out input_genotype_file_ped_format
$ gtool -P --ped input_genotype_file_ped_format.ped
--map input_genotype_file_ped_format.map
--og input_genotype_file_impute_format
--os input_genotype_file_sample
$ impute2 -g input_genotype_file_impute_format -m genetic_map_file
-int start end -o output_phased_haplotypes_file -phase
```

- **MaCH:**

```
$ mach1 -d input_genotype_mach_markers -p input_genotype_mach_genotypes
--round 50 --states 200 --phase -o output_phased_haplotypes_file
```

- **SHAPEIT2:**

```
$ shapeit2 -B input_genotype_file -M genetic_map_file
-O output_phased_haplotypes_file
```

- **SHAPEIT3:**

```
$ shapeit3.r884.1 --fast -B input_genotype_file -M genetic_map_file
-O output_phased_haplotypes_file
```

## B.7 Genotype imputation tool running

- **Beagle5:**

```
$ java -Xmx36g -jar beagle.16May19.351.jar
gt=input_phased_haplotype_file.vcf.gz ref=reference_panel_file.vcf.gz
map=genetic_map_file out=output_imputed_file
```

- **Minimac3:**

```
$ Minimac3 --refHaps reference_panel_file.vcf.gz
--haps input_phased_haplotype_file.vcf.gz
--chr phased_chr --prefix output_imputed_file
```

- **pbwt:**

```
$ pbwt -readVcfGT reference_panel_file.vcf.gz
-writeAll reference_panel_file_pbwt_format
$ pbwt -readVcfGT input_phased_haplotype_file.vcf.gz
-referenceImpute reference_panel_file_pbwt_format
-writeVcfGz output_imputed_file
```

# Bibliography


Abbas-Aghababazadeh, F., Li, Q., and Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PloS one*, 13(10):e0206312.

Albert, F. W. and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197.

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., and Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in x-linked mecp2, encoding methyl-cpg-binding protein 2. *Nature genetics*, 23(2):185.

Andrey, G. and Mundlos, S. (2017). The three-dimensional genome: regulating gene expression during pluripotency and development. *Development*, 144(20):3646–3658.

Anthony T., A. (2008). Dna packaging: Nucleosomes and chromatin. *Nature Education*, 1(26).

Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, Å., et al. (2003). Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, 72(5):1117–1130.

Balding, D. J., Bishop, M., and Cannings, C. (2008). *Handbook of statistical genetics*. John Wiley & Sons.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2004). Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265.

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nature genetics*, 36(5):431.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bezzina, C. R., Rook, M. B., Groenewegen, W. A., Herfst, L. J., van der Wal, A. C., Lam, J., Jongsma, H. J., Wilde, A. A., and Mannens, M. M. (2003). Compound heterozygosity for mutations (w156x and r225w) in scn5a associated with severe cardiac conduction disturbances and degenerative changes in the conduction system. *Circulation research*, 92(2):159–168.

Bonifer, C. and Cockerill, P. N. (2011). Chromatin mechanisms regulating gene expression in health and disease. In *Epigenetic Contributions in Autoimmune Disease*, pages 12–25. Springer.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C., and Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular ecology*, 13(11):3261–3273.

Brown, A. A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.-F., Richards, J. B., Small, K. S., Spector, T. D., Dermitzakis, E. T., and Durbin, R. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *Elife*, 3:e01381.

Brown, R., Kichaev, G., Mancuso, N., Boocock, J., and Pasaniuc, B. (2017). Enhanced methods to detect haplotypic effects on gene expression. *Bioinformatics*, 33(15):2307–2313.

Browning, B. L. and Browning, S. R. (2007a). Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(5):365–375.

Browning, B. L. and Browning, S. R. (2008). Haplotypic analysis of wellcome trust case control consortium data. *Human genetics*, 123(3):273–280.

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348.

Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics*, 124(5):439–450.

Browning, S. R. and Browning, B. L. (2007b). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714.

Buckland, P. R. (2004). Allele-specific gene expression differences in humans. *Human molecular genetics*, 13(suppl_2):R255–R260.

Buermans, H. and Den Dunnen, J. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1932–1941.

Bukowicki, M., Franssen, S. U., and Schlötterer, C. (2016). High rates of phasing errors in highly polymorphic species with low levels of linkage disequilibrium. *Molecular ecology resources*, 16(4):874–882.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2018). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012.

Camargo, A., Azuaje, F., Wang, H., and Zheng, H. (2008). Permutation–based statistical tests for multiple hypotheses. *Source code for biology and medicine*, 3(1):15.

Carmelo, V. A., Kogelman, L. J., Madsen, M. B., and Kadarmideen, H. N. (2018). Wish-r–a fast and efficient tool for construction of epistatic networks for complex traits and diseases. *BMC bioinformatics*, 19(1):277.

Carulli, J. P., Artinger, M., Swain, P. M., Root, C. D., Chee, L., Tulig, C., Guerin, J., Osborne, M., Stein, G., Lian, J., et al. (1998). High throughput analysis of differential gene expression. *Journal of Cellular Biochemistry*, 72(S30–31):286–296.

Chen, S.-Y., Feng, Z., and Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease*, 9(6):1725.

Chen, X., Shi, X., Xu, X., Wang, Z., Mills, R., Lee, C., and Xu, J. (2012). A two-graph guided multi-task lasso approach for eqtl mapping. In *Artificial Intelligence and Statistics*, pages 208–217.

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58.

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Montgomery, S. B., et al. (2017). The impact of structural variation on human gene expression. *Nature genetics*, 49(5):692.

Choi, Y., Chan, A. P., Kirkness, E., Telenti, A., and Schork, N. J. (2018). Comparison of phasing strategies for whole human genomes. *PLoS genetics*, 14(4):e1007308.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

Clark, A. G. (1990). Inference of haplotypes from pcr-amplified samples of diploid populations. *Molecular biology and evolution*, 7(2):111–122.

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998). A dna polymorphism discovery resource for research on human genetic variation. *Genome research*, 8(12):1229–1231.

Consortium, . G. P. et al. (2012a). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56.

Consortium, E. P. et al. (2012b). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Consortium, G. et al. (2015). The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Consortium, G. et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204.

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal, R., Lupien, M., Markowitz, S., Scacheri, P. C., et al. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome research*, 24(1):1–13.

Crawford, D. C. and Nickerson, D. A. (2005). Definition and clinical importance of haplotypes. *Annu. Rev. Med.*, 56:303–320.

Curtis, D. and Sham, P. C. (2006). Estimated haplotype counts from case-control samples cannot be treated as observed counts. *The American Journal of Human Genetics*, 78(4):729–731.

Das, S., Abecasis, G. R., and Browning, B. L. (2018). Genotype imputation from large reference panels. *Annual review of genomics and human genetics*, 19:73–96.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284.

De Boulle, K., Verkerk, A. J., Reyniers, E., Vits, L., Hendrickx, J., Van Roy, B., Van Den Bos, F., de Graaff, E., Oostra, B. A., and Willems, P. J. (1993). A point mutation in the fmr-1 gene associated with fragile x mental retardation. *Nature genetics*, 3(1):31.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181.

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature communications*, 10(1):1–10.

Duffy, D. L., Montgomery, G. W., Chen, W., Zhao, Z. Z., Le, L., James, M. R., Hayward, N. K., Martin, N. G., and Sturm, R. A. (2007). A three–single-nucleotide polymorphism haplotype in intron 1 of oca2 explains most human eye-color variation. *The American Journal of Human Genetics*, 80(2):241–252.

Durbin, R. (2014). Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics*, 30(9):1266–1272.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution*, 12(5):921–927.

Fadista, J., Manning, A. K., Florez, J. C., and Groop, L. (2016). The (in) famous gwas p-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8):1202.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–1318.

França, L. T., Carrilho, E., and Kist, T. B. (2002). A review of dna sequencing techniques. *Quarterly reviews of biophysics*, 35(2):169–200.

Francesconi, M. and Lehner, B. (2014). The effects of genetic variation on gene expression dynamics during development. *Nature*, 505(7482):208.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.

Gådin, J. R., Buil, A., Colantuoni, C., Jaffe, A. E., Nielsen, J., Shin, J.-H., Hyde, T. M., Kleinman, J. E., Plath, N., Eriksson, P., et al. (2019). Comparison of quantitative trait loci methods: Total expression and allelic imbalance method in brain rna-seq. *PloS one*, 14(6):e0217765.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091.

Garnier, S., Truong, V., Brocheton, J., Zeller, T., Rovital, M., Wild, P. S., Ziegler, A., Munzel, T., Tiret, L., Blankenberg, S., et al. (2013). Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. *PLoS genetics*, 9(1):e1003240.

Gilad, Y., Rifkin, S. A., and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in genetics*, 24(8):408–415.

Gohlke, J. M., Thomas, R., Zhang, Y., Rosenstein, M. C., Davis, A. P., Murphy, C., Becker, K. G., Mattingly, C. J., and Portier, C. J. (2009). Genetic and environmental pathways to complex diseases. *BMC Systems Biology*, 3(1):46.

Goudey, B., Abedini, M., Hopper, J. L., Inouye, M., Makalic, E., Schmidt, D. F., Wagner, J., Zhou, Z., Zobel, J., and Reumann, M. (2015). High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in genome wide association studies. *Health information science and systems*, 3(S1):S3.

Goudey, B., Rawlinson, D., Wang, Q., Shi, F., Ferra, H., Campbell, R. M., Stern, L., Inouye, M. T., Ong, C. S., and Kowalczyk, A. (2013). Gwis-model-free, fast and exhaustive search for epistatic interactions in case-control gwas. *BMC genomics*, 14(3):S10.

Goudey, B. W. (2016). *Detection of epistasis in genome-wide association studies*. PhD thesis, The University of Melbourne.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245.

Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8.

Heller, M. J. (2002). Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153.

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A. K., McRae, A. F., Yang, J., Gibson, G., Martin, N. G., Metspalu, A., et al. (2014). Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249.

Herzig, A. F., Nutile, T., Babron, M.-C., Ciullo, M., Bellenguez, C., and Leutenegger, A.-L. (2018). Strategies for phasing and imputation in a population isolate. *Genetic epidemiology*.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386.

Howard, D. M., Hall, L. S., Hafferty, J. D., Zeng, Y., Adams, M. J., Clarke, T.-K., Porteous, D. J., Nagy, R., Hayward, C., Smith, B. H., et al. (2017). Genome-wide haplotype-based association analysis of major depressive disorder in generation scotland and uk biobank. *Translational psychiatry*, 7(11):1263.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529.

Jackson, M., Marks, L., May, G. H., and Wilson, J. B. (2018). The genetic basis of disease. *Essays in biochemistry*, 62(5):643–723.

Jaffe, A. E., Straub, R. E., Shin, J. H., Tao, R., Gao, Y., Collado-Torres, L., Kam-Thong, T., Xi, H. S., Quan, J., Chen, Q., et al. (2018). Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci*, 21(8):1117–1125.

Jirtle, R. L. and Skinner, M. K. (2007). Environmental epigenomics and disease susceptibility. *Nature reviews genetics*, 8(4):253.

Jorde, L. B. and Wooding, S. P. (2004). Genetic variation, classification and'race'. *Nature genetics*, 36:S28–S33.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5).

Kenny, E. E., Gusev, A., Riegel, K., Lütjohann, D., Lowe, J. K., Salit, J., Maller, J. B., Stoffel, M., Daly, M. J., Altshuler, D. M., et al. (2009). Systematic haplotype analysis resolves a complex plasma plant sterol locus on the micronesian island of kosrae. *Proceedings of the National Academy of Sciences*, 106(33):13886–13891.

Kerem, B.-s., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., and Tsui, L.-C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080.

Kim, S. A., Cho, C.-S., Kim, S.-R., Bull, S. B., and Yoo, Y. J. (2018). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated snps. *Bioinformatics*, 34(3):388–397.

Klau, G. W. and Marschall, T. (2017). A guided tour to computational haplotyping. In *Conference on Computability in Europe*, pages 50–63. Springer.

Knight, J. C. (2004). Allele-specific gene expression uncovered. *Trends in Genetics*, 20(3):113–116.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068.

Kreimer, A. and Pe'er, I. (2013). Variants in exons and in transcription factors affect gene expression in trans. *Genome biology*, 14(7):R71.

Lambert, J.-C., Grenier-Boley, B., Harold, D., Zelenika, D., Chouraki, V., Kamatani, Y., Sleegers, K., Ikram, M., Hiltunen, M., Reitz, C., et al. (2013). Genome-wide haplotype association study identifies the frmd4a gene as a risk locus for alzheimer's disease. *Molecular psychiatry*, 18(4):461–470.

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506.

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., et al. (2010). Quality

control and quality assurance in genotypic data for genome-wide association studies. *Genetic epidemiology*, 34(6):591–602.

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS genetics*, 8(1):e1002453.

Leal, S. M., Yan, K., and Müller-Myhsok, B. (2005). Simped: a simulation program to generate haplotype and genotype data for pedigree structures. *Human heredity*, 60(2):119–122.

Lewontin, R. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics*, 49(1):49.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

Li, Y., Grupe, A., Rowland, C., Nowotny, P., Kauwe, J. S., Smemo, S., Hinrichs, A., Tacey, K., Toombs, T. A., Kwok, S., et al. (2006). Dapk1 variants are associated with alzheimer's disease and allele-specific expression. *Human molecular genetics*, 15(17):2560–2568.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics*, 10:387–406.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834.

Liang, P. and Pardee, A. B. (2003). Analysing differential gene expression in cancer. *Nature Reviews Cancer*, 3(11):869.

Lin, S., Chakravarti, A., and Cutler, D. J. (2004). Haplotype and missing data inference in nuclear families. *Genome Research*, 14(8):1624–1632.

Lin, S., Cutler, D. J., Zwick, M. E., and Chakravarti, A. (2002). Haplotype inference in random population samples. *The American Journal of Human Genetics*, 71(5):1129–1137.

Liu, N., Zhang, K., and Zhao, H. (2008). Haplotype-association analysis. *Advances in genetics*, 60:335–405.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., et al. (2016a). Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443.

Loh, P.-R., Palamara, P. F., and Price, A. L. (2016b). Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*, 48(7):811.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580.

Lv, H., Zhang, M., Shang, Z., Li, J., Zhang, S., Lian, D., and Zhang, R. (2017). Genome-wide haplotype association study identify the fgfr2 gene as a risk gene for acute myeloid leukemia. *Oncotarget*, 8(5):7891.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2016). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565.

Makałowski, W. (2001). The human genome structure and organization. *Acta Biochim. Pol*, 48:587–598.

Manolio, T. A. (2017). In retrospect: A decade of shared genomic associations. *Nature*, 546(7658):360–361.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *The American Journal of Human Genetics*, 78(3):437–450.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499.

Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279.

McCombie, W. R., McPherson, J. D., and Mardis, E. R. (2018). Next-generation sequencing technologies. *Cold Spring Harbor perspectives in medicine*, page a036798.

McDonald, J. H. (2009). *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD.

Miar, Y., Sargolzaei, M., and Schenkel, F. S. (2017). A comparison of different algorithms for phasing haplotypes using holstein cattle genotypes and pedigree data. *Journal of Dairy Science*, 100(4):2837–2849.

Miller, M. B. and Tang, Y.-W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical microbiology reviews*, 22(4):611–633.

Montana, G. (2005). Hapsim: a simulation tool for generating haplotype data with pre-specified allele frequencies and ld coefficients. *Bioinformatics*, 21(23):4309–4311.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773.

Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4).

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in genetics*, 6:285.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443.

Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 70(1):157–169.

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.-F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. *Nature genetics*, 48(7):817.

O'Connor, C. M., Adams, J. U., and Fairman, J. (2010). Essentials of cell biology. *Cambridge, MA: NPG Education*, 1.

Park, T., Yi, S.-G., Kang, S.-H., Lee, S., Lee, Y.-S., and Simon, R. (2003). Evaluation of normalization methods for microarray data. *BMC bioinformatics*, 4(1):33.

Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics*, 11(8):533.

Pattaro, C., Ruczinski, I., Fallin, D. M., and Parmigiani, G. (2008). Haplotype block partitioning as a tool for dimensionality reduction in snp association studies. *BMC genomics*, 9(1):405.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768.

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6(11):847.

Pook, T., Schlather, M., de Los Campos, G., Mayer, M., Schoen, C. C., and Simianer, H. (2019). Haploblocker: Creation of subgroup-specific haplotype blocks and libraries. *Genetics*, 212(4):1045–1061.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Qin, Z. S., Niu, T., and Liu, J. S. (2002). Partition-ligation–expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 71(5):1242–1247.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of computational biology*, 8(6):557–569.

Rubinacci, S., Delaneau, O., and Marchini, J. (2020). Genotype imputation using the positional burrows wheeler transform. *bioRxiv*, page 797944.

Ryckman, K. and Williams, S. M. (2008). Calculation and use of the hardy-weinberg model in association studies. *Current protocols in human genetics*, 57(1):1–18.

Saad, M. N., Mabrouk, M. S., Eldeib, A. M., and Shaker, O. G. (2018). Comparative study for haplotype block partitioning methods–evidence from chromosome 6 of the north american rheumatoid arthritis consortium (narac) dataset. *PloS one*, 13(12):e0209603.

Salem, R. M., Wessel, J., and Schork, N. J. (2005). A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Human Genomics*, 2(1):39.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, 15(11):1576–1583.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9):1748–1759.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470.

Sealfon, S. C. and Chu, T. T. (2011). Rna and dna microarrays. In *Biological Microarrays*, pages 3–34. Springer.

Shabalin, A. A. (2012). Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358.

Shang, Z., Lv, H., Zhang, M., Duan, L., Wang, S., Li, J., Liu, G., Ruijie, Z., and Jiang, Y. (2015). Genome-wide haplotype association study identify tnfrsf1a, casp7, lrp1b, cdh1 and tg genes associated with alzheimer's disease in caribbean hispanic individuals. *Oncotarget*, 6(40):42504.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345.

Shlyakhter, I., Sabeti, P. C., and Schaffner, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429.

Snyder, M. W., Adey, A., Kitzman, J. O., and Shendure, J. (2015). Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics*, 16(6):344.

Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human molecular genetics*, 24(R1):R111–R119.

Sprecher, E., Molho-Pessach, V., Ingber, A., Sagi, E., Indelman, M., and Bergman, R. (2004). Homozygous splice site mutations in pkp1 result in loss of epidermal

plakophilin 1 expression and underlie ectodermal dysplasia/skin fragility syndrome in two consanguineous families. *Journal of Investigative Dermatology*, 122(3):647–651.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500.

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5):1162–1169.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989.

Studer, R. A., Dessailly, B. H., and Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochemical Journal*, 449(3):581–594.

Stumpf, M. P. and McVean, G. A. (2003). Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, 4(12):959–968.

Sul, J. H., Raj, T., de Jong, S., De Bakker, P. I., Raychaudhuri, S., Ophoff, R. A., Stranger, B. E., Eskin, E., and Han, B. (2015). Accurate and fast multiple-testing correction in eqtl studies. *The American Journal of Human Genetics*, 96(6):857–868.

Taliun, D., Gamper, J., Leser, U., and Pattaro, C. (2015). Fast sampling-based whole-genome haplotype block recognition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(2):315–325.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., and Schork, N. J. (2011). The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215.

Tian, L., Quitadamo, A., Lin, F., and Shi, X. (2014). Methods for population-based eqtl analysis in human genetics. *Tsinghua Science and Technology*, 19(6):624–634.

Torres, A., Brownstein, C. A., Tembulkar, S. K., Graber, K., Genetti, C., Kleiman, R. J., Sweadner, K. J., Mavros, C., Liu, K. X., Smedemark-Margulies, N., et al. (2018). De novo atp1a3 and compound heterozygous nlrp3 mutations in a child with autism spectrum disorder, episodic fatigue and somnolence, and muckle-wells syndrome. *Molecular Genetics and Metabolism Reports*, 16:23–29.

Tregouet, D.-A. and Garelle, V. (2007). A new java interface implementation of thesias: testing haplotype effects in association studies. *Bioinformatics*, 23(8):1038–1039.

Trégouët, D.-A., König, I. R., Erdmann, J., Munteanu, A., Braund, P. S., Hall, A. S., Großhennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M., et al. (2009). Genome-wide haplotype association study identifies the slc22a3-lpal2-lpa gene cluster as a risk locus for coronary artery disease. *Nature genetics*, 41(3):283.

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., De Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., et al. (2011). Quality control procedures for genome-wide association studies. *Current protocols in human genetics*, 68(1):1–19.

Verma, S. S., De Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., Mukherjee, S., Jarvik, G. P., Kottyan, L. C., Burt, A., et al. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in genetics*, 5:370.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

Wall, J. D. and Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587.

Wang, J. (2018). Estimating genotyping errors from genotype and reconstructed pedigree data. *Methods in Ecology and Evolution*, 9(1):109–120.

Wang, L. and Xu, Y. (2003). Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–1780.

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of

population history, recombination, and mutation. *The American Journal of Human Genetics*, 71(5):1227–1234.

Wang, Q., Lv, H., Lv, W., Shi, M., Zhang, M., Luan, M., Zhu, H., Zhang, R., and Jiang, Y. (2015). Genome-wide haplotype association study identifies blm as a risk gene for prostate cancer in chinese population. *Tumor Biology*, 36(4):2703–2707.

Wang, S., Zhao, J. H., An, P., Guo, X., Jensen, R. A., Marten, J., Huffman, J. E., Meidtner, K., Boeing, H., Campbell, A., et al. (2016). General framework for meta-analysis of haplotype association tests. *Genetic epidemiology*, 40(3):244–252.

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with fuma. *Nature communications*, 8(1):1826.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B., and Bassett, D. E. (2006). Rosetta error model for gene expression analysis. *Bioinformatics*, 22(9):1111–1121.

Westra, H.-J., Martínez-Bonet, M., Onengut-Gumuscu, S., Lee, A., Luo, Y., Teslovich, N., Worthington, J., Martin, J., Huizinga, T., Klareskog, L., et al. (2018). Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat. Genet*, 50:1366–1374.

Wilkie, A. O., Slaney, S. F., Oldridge, M., Poole, M. D., Ashworth, G. J., Hockley, A. D., Hayward, R. D., David, D. J., Pulleyn, L. J., Rutland, P., et al. (1995). Apert syndrome results from localized mutations of fgfr2 and is allelic with crouzon syndrome. *Nature genetics*, 9(2):165.

Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H., and Reich, D. (2012). Phasing of many thousands of genotyped samples. *The American Journal of Human Genetics*, 91(2):238–251.

Wu, Y., Fan, H., Wang, Y., Zhang, L., Gao, X., Chen, Y., Li, J., Ren, H., and Gao, H. (2014). Genome-wide association studies using haplotypes and individual snps in simmental cattle. *PLoS One*, 9(10):e109330.

Yang, Y., Li, S. S., Chien, J. W., Andriesen, J., and Zhao, L. P. (2008). A systematic search for snps/haplotypes associated with disease phenotypes using a haplotype-based stepwise procedure. *BMC genetics*, 9(1):90.

Yap, C. X., Lloyd-Jones, L., Holloway, A., Smartt, P., Wray, N. R., Gratten, J., and Powell, J. E. (2018). Trans-eqtls identified in whole blood have limited influence on complex disease biology. *European Journal of Human Genetics*, 26(9):1361.

Ying, D., Li, M. J., Sham, P. C., and Li, M. (2018). A powerful approach reveals numerous expression quantitative trait haplotypes in multiple tissues. *Bioinformatics*, 34(18):3145–3150.

Zhang, F. and Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1):R102–R110.

Zhang, K., Calabrese, P., Nordborg, M., and Sun, F. (2002). Haplotype block structure and its applications to association studies: power and study designs. *The American Journal of Human Genetics*, 71(6):1386–1394.

Zhong, K., Zhu, G., Jing, X., Hendriks, A. E. J., Drop, S. L., Ikram, M. A., Gordon, S., Zeng, C., Uitterlinden, A. G., Martin, N. G., et al. (2017). Genome-wide compound heterozygote analysis highlights alleles associated with adult height in europeans. *Human genetics*, 136(11-12):1407–1417.

Zhu, X., Zhang, S., Kan, D., and Cooper, R. (2003). Haplotype block definition and its application. In *Biocomputing 2004*, pages 152–163. World Scientific.