# Next Generation of Genotype Imputation Methods

by

Sayantan Das

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2017

Doctoral Committee:

Professor Gonçalo Abecasis, Chair
Professor Michael Lee Boehnke
Assistant Professor Hyun Min Kang
Professor Bhramar Mukherjee
Assistant Professor Jenna Wiens

Sayantan Das

sayantan@umich.edu

ORCID iD: 0000-0001-6346-1590

# DEDICATION

This dissertation is dedicated to my parents, family and friends.

# ACKNOWLEDGMENTS

I would first like to thank my advisor Gonçalo Abecasis for sharing his rare insights on genetics research and for steering me in the right direction whenever I needed it. I would also like to thank the other members of my dissertation committee, Hyun M. Kang and Michael Boehnke for their continued support and guidance over the last 6 years, Bhramar Mukherjee for making me feel like a part of family even while miles away from home, and Jenna Wiens for her advice and assistance with this thesis.

I am grateful to Lam C. Tsoi for being a valuable mentor during the initial years of graduate school, Alan Kwong for always being there as a great friend and a valuable colleague, and Gregory Zajac, Chris Harvey, and Amanda Artis for lending their expertise with the language of the material. My acknowledgements should also reach out to the members of the Haplotype Reference Consortium and the Steering Committee for Trans-Omics for Precision Medicine for providing me with the remarkable opportunity to analyze their data.

To my many friends, you should know that your support and encouragement was worth more than I can express on paper; Ankita and Rashmi, for teaching me the importance of passion, Debsmita, for compelling me to get a PhD and being so supportive and understanding, Jessica, for being such a great friend and making Ann Arbor feel like home, and Avik, for being an excellent accomplice since college.

Finally, I must express my very profound gratitude to my parents and my sister for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

# TABLE OF CONTENTS

**Dedication** . . . . . **ii**

**Acknowledgments** . . . . . **iii**

**List of Figures** . . . . . **vi**

**List of Tables** . . . . . **viii**

**List of Appendices** . . . . . **ix**

**List of Abbreviations** . . . . . **x**

**Abstract** . . . . . **xii**

**Chapter**

**1 Introduction** . . . . . **1**

1.1 Genotype Imputation . . . . . 3
1.1.1 Development of Imputation Tools . . . . . 6
1.1.2 Development of Reference Panels . . . . . 11
1.1.3 Improving Imputation Quality . . . . . 12
1.2 Outline of Dissertation . . . . . 15

**2 Haplotype Reference Consortium - A Reference Panel of 64,976 Haplotypes for Genotype Imputation** . . . . . **17**

2.1 Introduction . . . . . 17
2.2 Methods . . . . . 17
2.2.1 Quality Control . . . . . 18
2.2.2 Phasing . . . . . 18
2.3 Results . . . . . 19
2.3.1 Improvement in Imputation Accuracy . . . . . 20
2.3.2 Higher Resolution for Fine Mapping . . . . . 20
2.4 Discussion . . . . . 23

**3 Minimac3 : Next-generation Genotype Imputation Service and Methods** . . . . . **24**

3.1 Introduction . . . . . 24
3.2 Methods . . . . . 25

3.2.1 State Space Reduction . . . 27
3.2.2 Computational Complexity and Optimal Allocation . . . 29
3.2.3 Comparison of Contemporary Tools . . . 30
3.3 Results . . . 31
3.3.1 Faster Imputation . . . 31
3.3.2 Scaling with Large Panels . . . 32
3.3.3 Compressed File Structure . . . 35
3.3.4 Imputation as a Service . . . 36
3.4 Discussion . . . 36

**4 Minimac4 - Faster Imputation through Aggressive State Space Reduction of Hidden Markov Models . . . 38**

4.1 Introduction . . . 38
4.2 Methods . . . 39
4.2.1 Aggressive State Space Reduction . . . 39
4.2.2 Imputation of Ungenotyped Markers . . . 43
4.2.3 Software Engineering Techniques . . . 45
4.3 Results . . . 47
4.3.1 Comparison of Imputation Tools . . . 47
4.3.2 Comparison of Imputation Accuracy . . . 48
4.3.3 Comparison of Reference Panels . . . 48
4.3.4 Sensitivity to Parameter Values . . . 51
4.3.5 Imputation in Non-European Samples . . . 54
4.4 Discussion . . . 54

**5 MetaMinimac - A Simple and Flexible Method to Combine Imputed Data from Multiple Reference Panels . . . 57**

5.1 Introduction . . . 57
5.1.1 Motivation . . . 57
5.2 Methods . . . 60
5.2.1 Model Description . . . 60
5.2.2 Leave-One-Out Imputation . . . 60
5.2.3 Estimation of Weights . . . 64
5.3 Results . . . 64
5.3.1 Meta-imputing Samples of Admixed Ancestry . . . 64
5.3.2 Meta-imputation of Three Reference Panels . . . 66
5.3.3 Additional Benefit . . . 66
5.4 Discussion . . . 67

**6 Conclusion . . . 69**

**Appendices . . . 72**

**Bibliography . . . 106**

# LIST OF FIGURES


1.1 A brief time-line summarizing the major developments in genotype imputation . . . . 7
1.2 An illustration of genotype imputation . . . . . 8

2.1 Effect of variant filtering on Ts/Tv ratio in the HRC panel . . . . . 19
2.2 Improved imputation accuracy from the HRC panel . . . . . 21
2.3 Improved association signal from the HRC panel . . . . . 22

3.1 Overview of state space reduction in minimac3 . . . . . 26

4.1 Overview of aggressive state space reduction in minimac4 . . . . . 41
4.2 Assembly of flanking regions for imputation . . . . . 44
4.3 Comparison of 1000G Phase 1, Phase 3, HRC, and TOPMed . . . . . 52
4.4 Effect of chunking on imputation accuracy . . . . . 53

5.1 Imputation into 20 Norwegian samples . . . . . 59
5.2 Schematic diagram explaining leave-one-out imputation . . . . . 62
5.3 Meta-imputation in African American samples . . . . . 65

6.1 Summary of development from MaCH to Minimac4 . . . . . 70
6.2 Cost savings in imputation and improvement since 2010 . . . . . 70

A.1 Illustration of hidden Markov model . . . . . 72

C.1 Illustration of site filtering strategy . . . . . 83

E.1 Posterior probability of reference haplotypes . . . . . 93

F.1 Ancestry of 1,000 TOPMed samples . . . . . 95

G.1 Meta-imputation in samples of mixed ancestry . . . . . 97
G.2 Meta-imputation of 3 reference panels . . . . . 98
G.3 Plot of template switch probability across the chromosome . . . . . 100

H.1 Imputation accuracy in African American samples . . . . . 101
H.2 Imputation accuracy in admixed American samples . . . . . 102
H.3 Imputation accuracy in East Asian samples . . . . . 103
H.4 Imputation accuracy in South Asian samples . . . . . 104

I.1 Geo-location of users of Michigan imputation server . . . . . 105

I.2 Tweets by users of Michigan imputation server . . . . . 105

# LIST OF TABLES


1.1 Summary of imputation tools that employ HMM . . . . . 10

3.1 Description of reference panels used for imputation experiments . . . . . 31
3.2 Imputation accuracy based on simulated haplotypes . . . . . 32
3.3 Computational requirement of minimac3 and other imputation tools . . . . . 33
3.4 Imputation accuracy of minimac3 and other imputation tools . . . . . 34

4.2 Computational requirements of minimac4 and other imputation tools . . . . . 49
4.1 Comparison of imputation time for different reference panels . . . . . 50
4.3 Predicted run-time for different reference panels . . . . . 51

C.1 Evaluation of genotype calling process . . . . . 80
C.2 Summary of studies contributing samples to the HRC panel . . . . . 81
C.3 Details of duplicate removal in HRC . . . . . 82

D.1 Summary of data compression for different ancestries . . . . . 87
D.2 Comparison of VCF and m3vcf file format . . . . . 88

E.1 Imputation accuracy of minimac4 compared to minimac3 . . . . . 94

G.1 Imputation accuracy details for meta-imputation . . . . . 99

# LIST OF APPENDICES

**A A brief review of hidden Markov models** . . . . . **72**

**B Asymptotic results for association test with imputed genotypes** . . . . . **75**

**C Supplementary information for the Haplotype Reference Consortium** . . . . . **80**

**D Proof of methods and supplementary information for minimac3** . . . . . **84**

**E Proof of methods and supplementary diagrams for minimac4** . . . . . **89**

**F Principal components of 1,000 randomly selected TOPMed samples** . . . . . **95**

**G Supplementary tables/figures for meta-imputation** . . . . . **96**

**H Imputation accuracy graphs for samples of non-European ancestry** . . . . . **101**

**I Supplementary figures for the Michigan imputation server** . . . . . **105**

# LIST OF ABBREVIATIONS

**1000G** 1000 Genomes

**AF** allele frequency

**AFR** African

**ASW** African-American samples in Southwest USA

**BAM** binary alignment map

**CAAPA** Consortium on Asthma among African-ancestry Populations in the Americas

**CG** Complete Genomics

**EAS** East Asian

**EM** Expectation-Maximization

**EUR** European

**GWA** genome wide association

**HLA** human leukocyte antigen

**HMM** Hidden Markov Model

**HRC** Haplotype Reference Consortium

**LD** linkage disequilibrium

**loo** leave-one-out

**MAC** minor allele count

**MAF** minor allele frequency

**MCMC** Monte-Carlo Markov Chain

**OpenMP** Open Multi-Processing

**PBWT** Positional Burrows-Wheeler Transformation

**QC** quality control

**SNP** single nucleotide polymorphisms

**TOPMed** Trans-Omics for Precision Medicine

**VCF** Variant Call Format

**WGS** whole genome sequencing

**WTCCC** Wellcome Trust Case Control Consortium

# ABSTRACT

In the past several years, we have witnessed numerous human genetic studies that have systematically evaluated the contribution of genetic polymorphisms to various complex diseases, and enabled the evolution of multiple treatment strategies, particularly pharmaceutical therapies. Genotype imputation has been a key step in such studies - increasing the power of gene mapping analyses, facilitating harmonization of results across studies, and accelerating fine-mapping efforts. Imputation requires access to a reference panel of densely sequenced genomes and is a computationally intensive process, even with modern high performance computing. Furthermore, reference panels often have data privacy issues that inhibit users from having direct access to the data. The goal of this dissertation is to design novel strategies to address these challenges for the next generation of imputation methods.

In the first project, I describe our efforts to create a reference panel of ~$32,000$ individuals with ~$40$M variants by combining genetic information obtained across 20 whole genome sequencing studies (Haplotype Reference Consortium). In the second project, I describe a novel idea called 'state space reduction' that reduces computational requirements of genotype imputation by orders of magnitude without any loss of accuracy (minimac3). I also present a web-based platform for imputation that greatly improves user experience and productivity. In the third project, I extend the idea of state space reduction by implementing a more complex version of the strategy that produces additional cost savings (minimac4). In the fourth project, I introduce the idea of meta-imputation: a novel approach that integrates imputed data from multiple reference panels at overlapping sites without interfering in the imputation algorithm (MetaMinimac).

In summary, the purpose of this dissertation research is to develop statistical methods and com-

putational tools that will benefit other researchers in the next generation of human gene mapping studies. These imputation tools will detect rare variants with higher accuracy, consequently increasing the power of association studies.

# CHAPTER 1

## Introduction

The field of human genetics has made substantial advances since its genesis with studying simple Mendelian disorders a century ago. In the past decade, more attention has been paid to susceptibility to complex diseases[1] due to both genetic and environmental factors. In 2005, the first ever genome wide association (GWA)[2] study was conducted on unrelated patients with age-related macular degeneration (Klein et al. 2005). Although they had a very small sample (~100 cases and ~50 controls), it was the first study to find regulatory genes in inflammation pathways associated with macular degeneration, which was an unexpected and novel finding. In 2007, the Wellcome Trust Case Control Consortium (WTCCC) launched one of the largest GWA studies of its time, analyzing ~17,000 cases and controls for seven common diseases: bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I diabetes, and type II diabetes (WTCCC 2007). Along with replicating many previously implicated genetic loci, the study revealed multiple new risk loci for several of the diseases, including Crohn's disease and rheumatoid arthritis.

The success of these studies led to a rapid explosion of interest in the field; numerous GWA studies have systematically evaluated the contributions of genetic factors to various complex diseases (McCarthy et al. 2008) along with quantitative traits such as human height (Lango Allen et al. 2010), body mass index (Locke et al. 2015) and cholesterol (Global Lipids et al. 2013). These studies helped researchers reveal unexpected pathways in disease etiology, such as the importance of complement factor genes in macular degeneration (Klein et al. 2005), the role of the central

[1] Diseases that are influenced by a combination of multiple genes and environmental factors

[2] A study that tests for disease susceptibility across a genome-wide set of common genetic variants

nervous system in obesity susceptibility (Locke et al. 2015), and the function of genes in the autophagy pathway in Crohn's disease (Barrett et al. 2008). They have also provided evidence for previously suspected molecular mechanisms (e.g. the role of *IL-23* signaling in psoriasis by Nair et al. 2009 or the role of *APOE* in Alzheimer's disease by Coon et al. 2007) and enabled the development of new drugs and treatment strategies (Okada et al. 2013; Zhang et al. 2015). Although GWA studies now seem quite routine, they have greatly changed human genetics in the last 10 years by presenting a new systematic method of providing deeper insights about disease biology.

Despite their immense success, the magnitude of associations from GWA studies have been smaller than expected, and they have been able to explain only a limited proportion of disease heritability[3] (Manolio et al. 2009; Witte et al. 2014). One plausible explanation is that GWA studies use commercial genotyping arrays, which mainly analyze regions of common variation in the human genome. However, rare variants[4], which are more often associated with dramatic functional consequences, might play an important role in the missing heritability (Gibson 2012). A standard method to analyze rare variants is whole genome sequencing (WGS)[5]. Given sufficient coverage ($>$ 100x), WGS can detect the rarest of mutations, including somatic mutations, with very high accuracy (Alioto et al. 2015). Although next-generation technologies have significantly reduced the cost of sequencing a genome, it remains prohibitively expensive for a study to whole genome sequence a large number of samples (Goodwin et al. 2016). An alternative approach is to use custom designed genotyping arrays enriched with rarer variants by utilizing information from whole genome/exome sequencing projects (Hoffmann et al. 2011). Although more affordable, these custom arrays can only analyze variants that are already known to researchers and thus would not include the novel rare variants unique to the samples being assayed.

[3]Degree to which genetic variation accounts for disease susceptibility

[4] Variants with minor allele frequency less than 1%

[5]Process of determining the complete DNA sequence of an organism

## 1.1 Genotype Imputation - Sequencing on the Cheap

A more cost efficient way of analyzing rare variants is to impute them in a typical GWA study (Li et al. 2009b). Rare variants, although not directly assayed on GWA arrays, can be imputed using genetic information from a reference set of densely sequenced genomes. This method of estimating genotypes or genotype probabilities[6] at markers that have not been directly genotyped in a typical genetic study is known as 'genotype imputation'. One of the first studies to use genotype imputation was on type 2 diabetes in Finnish samples (Scott et al. 2007). Imputation helped the researchers identify and replicate multiple risk variants as well as compare their results to two other concurrent studies which used different genotyping arrays. Since then, imputation has been a key step in the analysis of human genetic studies - accelerating fine-mapping efforts, aiding the combination of results across studies (meta-analyses), and increasing the power of gene mapping analyses (Marchini et al. 2010). Some examples of the benefits of genotype imputation are given below.

**Fine-mapping.** Imputation provides a higher resolution view of a genetic region by adding more variants, thereby increasing the chances of identifying a causal variant. For example, a study on blood triglyceride levels that aimed to fine-map the *GCKR* gene found that the strongest signal came from a missense variant that was imputed and later confirmed by direct genotyping (Orho-Melander et al. 2008). Similarly, another study on the WTCCC data found the strongest signal in a type 2 diabetes gene (*TCF7L2*) coming from an imputed variant (rs7903146) instead of a genotyped variant (Marchini et al. 2007, Figure 4).

**Meta-analysis.** Imputation also helps in meta-analysis by facilitating the combination of results across studies. Different studies often use different genotyping arrays containing different sets of variants. For example, the Affymetrix® 6.0 SNP array includes only a portion of the SNPs

[6] probability of each possible genotype at each marker, which is beneficial if the true genotype cannot be imputed with high accuracy

included on the Illumina® 660K array. Genotype imputation can be used to generate a common set of variants which can then be analyzed across all the studies to boost power. This approach has been successful in identifying several new loci for different traits such as type 1 diabetes (Cooper et al. 2008) and psoriasis (Tsoi et al. 2015).

**Increasing Power of Association Studies.** Another benefit of imputation lies in increasing the power to detect an association signal. Although imputed genotypes are by definition correlated to the original genotypes, when single nucleotide polymorphisms (SNP) are only genotyped in a portion of the samples, imputation can increase the effective sample size by filling in the missing samples. This was demonstrated in a study on triglycerides and cholesterol, where a common variant in a known risk gene (*LDLR*) was missed when only the genotyped SNPs were analyzed but was then identified following imputation (Willer et al. 2008). This was because the genotyping chip used to assay the majority of samples not only lacked this particular SNP, the directly genotyped SNPs also inadequately tagged this SNP. Some simulation studies have shown that imputation can increase power by up to 10% when compared to testing only genotyped SNPs (Spencer et al. 2009, Table 1) while some others have predicted more modest gains (Anderson et al. 2008, Figure 3; Hao et al. 2009, Figure 3).

**Other benefits.** Imputed data have also been used to test for pleiotropic effects by imputing a known risk variant into multiple disease studies. For example, Hoffmann et al. 2015 found evidence suggesting a pleiotropic effect from a *HOXB13* mutation across multiple cancers. Imputation has been used to estimate other types of genetic variations such as copy number variants (Handsaker et al. 2015), classical human leukocyte antigen (HLA) alleles (Leslie et al. 2008), and amino acid polymorphisms that are difficult to genotype or sequence (Jia et al. 2013).

Genotype imputation has become an integral part of human genetic studies. Trait mapping using imputed variants is significantly cheaper than analyzing variants from whole genome sequencing. The only caveat lies in the loss of power. Testing for association using an imputed

variant on $N$ samples yields the same power as testing using the unobserved true variant on $Nr^2$ samples (where $r^2$ is the Pearson correlation between the true variant and the imputed variant). In other words, the effective sample size is reduced by $r^2$. Nevertheless, the effect size estimates from imputed data remain unbiased in a linear regression, provided the imputed variant is "well-calibrated[7]". More details and proofs are provided Appendix B. Although imputed data doesn't attenuate the effect sizes, the assumption of well-calibration is quite to possible to violate, leading to imperfect imputation results and consequently spurious association signals. This is illustrated by an extreme example where cases and controls are genotyped on two different platforms resulting in a marker of interest being genotyped in the cases but not in the controls. If a marker of interest is not well predicted by flanking markers, imputation will suggest the genotype distribution of the reference panel at that marker, and this can be a very poor assumption. Similarly, a genetically dissimilar reference panel can potentially increase the noise in imputation estimates, thereby leading to spurious or no association. In both these examples, the assumption of well-calibration is violated as the imputed dosage does not correctly represent the alternate allelic probability for the samples being analyzed.

In this dissertation, I present a series of four chapters that address some of the prevailing challenges in next generation imputation techniques. In Chapter 2, we describe our efforts to create a reference panel of ~32,000 individuals as a resource for genotype imputation (McCarthy et al. 2016). In Chapter 3, we describe a novel technique that significantly reduces the computational complexity of imputation algorithms by orders of magnitude without any loss of accuracy (Das et al. 2016). We also present a web-based platform that greatly facilitates imputation analysis and improves user experience and productivity. In Chapter 4, we describe methods that produce additional cost savings on imputation algorithms, thereby enabling users to analyze future studies with millions of samples at a negligible loss in accuracy. In Chapter 5, we introduce the idea of

[7] An imputed variant is said to be well-calibrated for a sample haplotype if the probability of observing the alternate allele for that haplotype is equal to its imputed alternate allele dosage, that is $P(G = 1|D) = D$ where $G$ is the observed allele at the haplotype ($G = 0, 1$) and $D$ is the imputed dosage ($0 \leq D \leq 1$)

meta-imputation, a novel approach that integrates imputed data from multiple reference panels at overlapping sites without interfering in the imputation algorithm.

To motivate the work described in this dissertation, we will first briefly review the development of genotype imputation tools in the last decade. Next, we will review different reference panels used in contemporary imputation analyses. Some major developments in the field of genotype imputation (including reference panels and imputation tools) are summarized in Figure 1.1, p.7. Finally, we will present an outline of the four chapters in this dissertation.

### 1.1.1 Development of Genotype Imputation Tools

Estimation of missing data has been a ubiquitous problem in most applications of statistics and human genetic studies are no exception. While this challenge could have been mostly solved by traditional statistical imputation methods, the advent of GWA studies ushered in a new era of imputation problems that mainstream methods were fundamentally incapable of solving (Yu et al. 2007). The primary reason is the extremely high rate of missingness; commercial GWA arrays only genotype ~0.1% of the human genome[8] while the remaining ~99.9% of the genome is missing and needs to be imputed. The second reason is that common statistical imputation techniques (linear regression, regression trees, k-nearest neighbors) lacked the sophistication to take advantage of genetic architecture (linkage patterns, recombination hot-spots, mutations, genotyping errors, etc.). These challenges unique to statistical genetics necessitated the development of computational tools created explicitly for genotype imputation in human gene studies.

**Intuition.** The basic intuition behind genotype imputation is as follows: individuals of similar ancestry, even if apparently unrelated, can be expected to share short stretches of chromosomal segments. Consequently, once a study sample is genotyped on a commercial array (with mostly missing data), the alleles at the genotyped markers can be used to match these short DNA segments between the study sample and a reference panel of densely sequenced genomes (with no missing

[8]The densest genotyping arrays cover ~5M variants while the human genome has ~3 billion variants

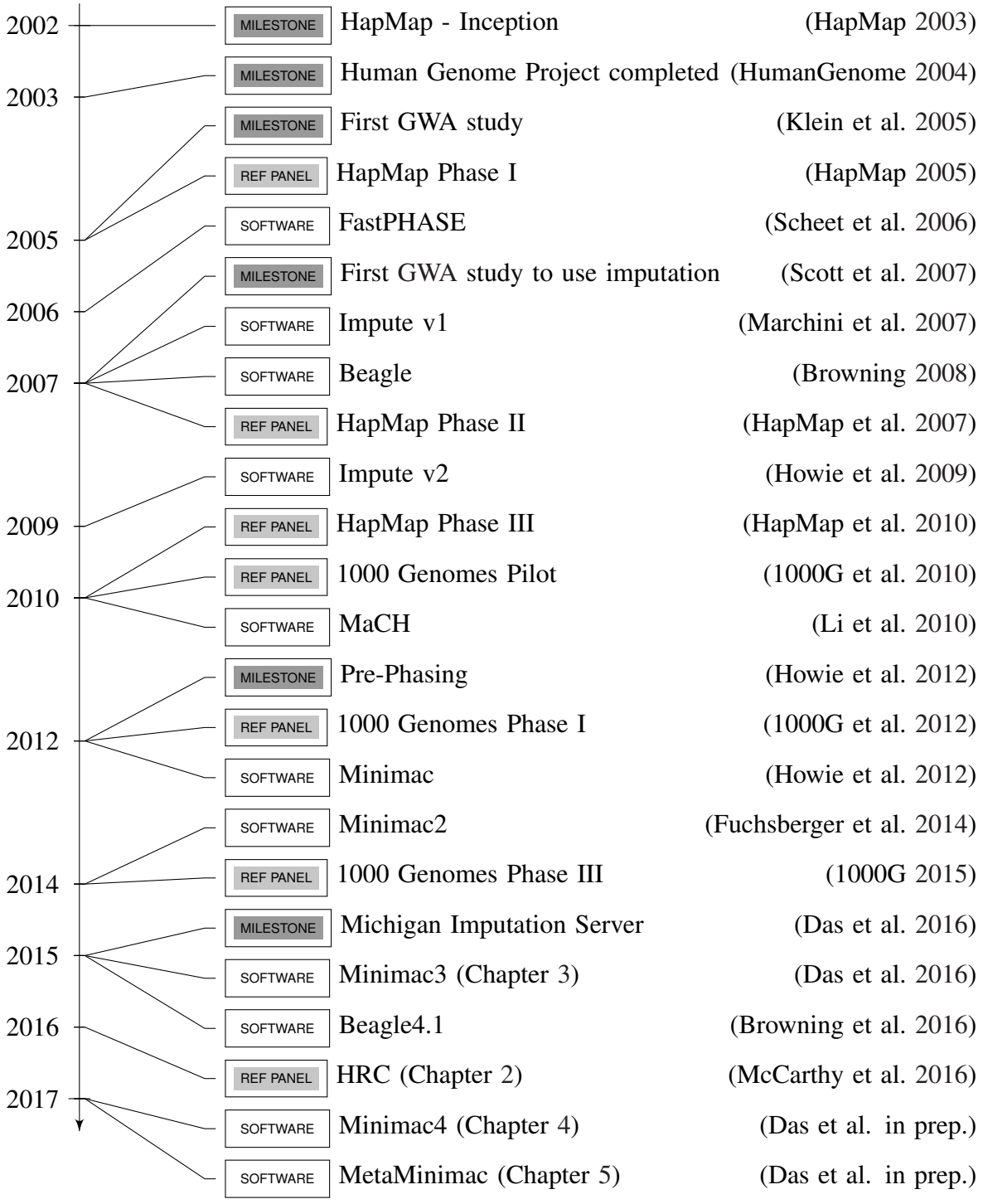


Figure 1.1: A brief time-line summarizing the major developments in genotype imputation. Each major development has been categorized as either a milestone (labels in dark gray), reference panel (light gray) or software (no shading)

data). In other words, a study haplotype can be represented as an imperfect mosaic of short segments of haplotypes (templates) from the reference panel, enabling one to impute the sites that were not genotyped. See Figure 1.2 for a simplified illustration. A probabilistic framework is needed to find estimated template switch positions and account for events like multiple matches and duplicate matches, a common occurrence in real applications.

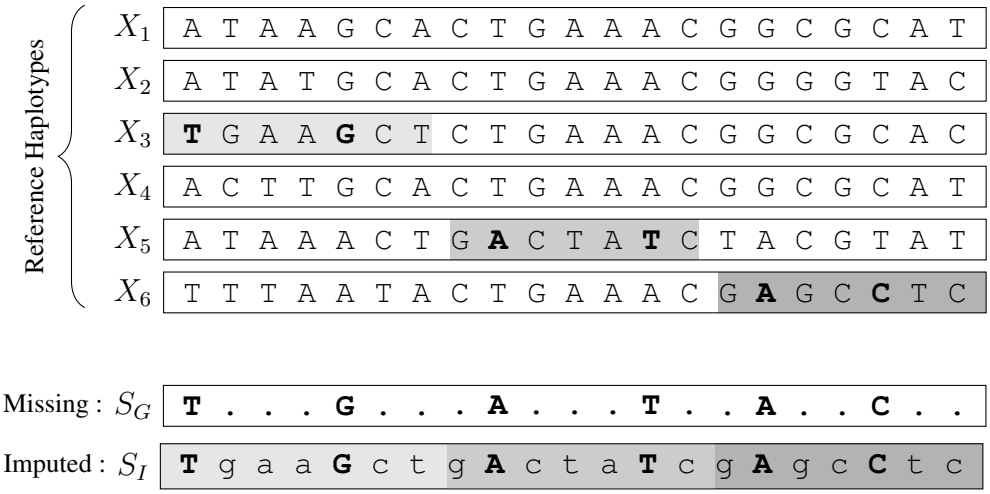


Figure 1.2: An illustration of genotype imputation. The figure explains the process of imputation for a study haplotype ($S_G$) genotyped at 6 markers (in bold) using a reference panel of sequenced haplotypes at 21 markers. The alleles in $S_G$ are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype $X_3$. In the second segment, the alleles A and T imply that haplotype $X_5$ might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes $X_3$, $X_5$, and $X_6$. Consequently, the missing sites can be imputed to obtain the final imputed haplotype $S_I$.

**Li and Stephens Model.** Although multiple research groups have developed numerous imputation tools over the last decade, the basic framework behind most of them is fundamentally the same and is known as the Li and Stephens model. It was first described by Li et al. 2003 and then later implemented as fastPHASE by Scheet et al. 2006. The framework uses a Hidden Markov Model (HMM)[9] to describe the data, where the observed genotypes of the study sample (with missing data) represent the observed states of the HMM while a set of reference haplotypes (with

[9] A statistical model in which the system being modeled is assumed to be a Markov chain with unobserved/hidden states. See Appendix A for more details

no missing data) represent the hidden states of the HMM (see Appendix A for a brief review of HMM). The HMM framework was immediately useful because it naturally represented sample genotype data as a mosaic of templates, and the points where the templates switch represented historical recombination events. The transition function between consecutive sites is a function of the linkage and recombination rates between the sites, while the emission probabilities are a function of mutation rates and genotyping error rates.

Although most contemporary imputation tools employ the HMM framework for their main algorithm, they differ from each other in how they define the state space and the parameters of the HMM. A brief summary of the major imputation tools in the last decade is given in Table 1.1, p.10. While fastPHASE, MaCH, and Impute were quite similar, the first version of Beagle was distinct from others because it did not employ the usual parameter estimates (transition and emission functions), and the haplotype model was constructed from the study samples as opposed to a reference panel (Browning 2008). However, the most recent version of Beagle has an updated imputation engine which uses the Li and Stephens model for imputation and is much more similar to the remaining tools (Browning et al. 2016).

Other imputation engines have employed direct methods of haplotype matching that don't employ the HMM framework. For example, PBWT uses the positional Burrows-Wheeler transformation to find 'set-maximal' matches at each marker of the genotyped sample, which are in turn used to assign alleles at the missing markers (Durbin 2014). While such prefix array methods are computationally highly efficient, their imputation accuracy is reduced because they cannot model historical recombination rates (see Figure 2.2, p.21). The earlier version of Beagle was essentially derived from a dynamic truncation of the positional prefix array, but it also employed a probabilistic model which reduced imputation speed. Additionally, some other methods (for example, PLINK (Purcell et al. 2007), SNPMStat (Lin et al. 2008), TUNA (Nicolae 2006), and UNPHASED (Dudbridge 2008)) use SNP tagging approaches to carry out imputation. Although these methods are quicker and simpler, they don't utilize information from the entire chromosome to perform imputation and hence generally provide less accurate estimates.

| Software | HMM State Space | HMM Parameter Functions |
|---|---|---|
| fastPHASE | All genotype configurations from a fixed number of localized haplotype clusters (say K) | Depends on recombination and mutation rates. Parameters fit using EM algorithm. |
| MaCH | All genotype configurations from all the reference haplotypes | Depends on recombination rate, mutation rate and genotyping error. Parameters fit using MCMC or EM algorithm |
| Impute | All genotype configurations from all the reference haplotypes | Depends on fine scale recombination map that is fixed and provided internally by the program |
| Beagle | All genotype configurations from a variable number of localized haplotye clusters | Empirical model with no explicit parameter functions |
| Minimac/2 | All possible reference haplotypes | Same as MaCH |
| Impute2 | All possible reference haplotypes | Same as Impute |
| Minimac3 | All unique allele sequences observed in reference data in a small genomic segment | Same as MaCH but parameters estimates are pre-calculated and fixed. |
| Beagle4.1 | All unique allele sequences observed in reference and target data at each "aggregate genotyped marker" | Depends on recombination rates and error rates which are fixed and pre-calculated |
| Minimac4 | Collapsed allele sequences from reference data that match at genotyped positions in small genomic segments | Same as minimac3. |

Table 1.1: Summary of imputation tools that employ HMM. This table describes the typical state space and parameter functions used to model the Li and Stephens framework (Li et al. 2003). Minimac and Impute2 were the first tools to use the 'pre-phasing' approach (Howie et al. 2012). Minimac3 and Beagle4.1 exploit local haplotype redundancy to reduce the size of the state space and hence the computational burden (Das et al. 2016; Browning et al. 2016).

**Pre-Phasing.** One of the major milestones to reduce computational burden in the Li and Stephens framework has been achieved in the approach of pre-phasing. This idea involves a two step-imputation process: the initial step of pre-phasing (i.e. haplotype estimation) of the GWA genotypes and a subsequent step of imputation into the estimated study haplotypes (Howie et al. 2012). These separate steps benefited researchers in two ways. First, the decomposed haplotypes could be re-used for imputation multiple times. Second, it reduced the complexity of the imputation step from quadratic to linear: this is because pre-phasing allowed matches to be found by comparison against phased haplotypes rather than comparison against all pairs of haplotypes. Earlier analytical methods integrated the imputation algorithm over phase uncertainty, thereby yielding marginally improved imputation accuracy in certain populations such as African Americans, (Howie et al. 2012). However, the ability to use the new method with larger reference panels mitigated the effect of that reduction. Minimac, Impute2, and Beagle4.1 are respective successors of MaCH, Impute, and Beagle that employ this pre-phasing approach.

Other major algorithmic developments and software engineering techniques helped to further improve computational efficiency. For example, minimac2 improved vector and matrix operations using a highly optimized `c++` library to increase its speed threefold over minimac (Fuchsberger et al. 2014). Similarly, Beagle4.1 employed parallelization, compact representation of reference haplotypes and genotype probabilities, and other improvements in software architecture to improve speed over their previous version (Browning et al. 2016). In Chapters 3 and 4 we describe our efforts to build even faster imputation tools.

### 1.1.2 Development of Reference Panels

Over the years, the quality of genotype imputation has benefitted greatly from improved genotyping technologies (e.g. high-density genotyping arrays) and more efficient analytical methods (e.g. pre-phasing), but most notably from the synthesis of genetic information through publicly available datasets. Examples of such datasets include the HapMap Project, the 1000 Genomes (1000G) Project, UK10K, the Haplotype Reference Consortium (HRC), and the Trans-Omics for

Precision Medicine (TOPMed) program (HapMap et al. 2010; 1000G 2015; UK10K et al. 2015; McCarthy et al. 2016; NHLBI 2015). However, the development of next generation sequencing technologies has led to a rapid increase in the size of public datasets used as reference panels for genotype imputation. For example, while the first release of the HapMap Project has an initial sample of approximately 250 sequenced individuals, phase 3 of 1000G Project had around 2,500 individuals, and upcoming reference panels from the TOPMed program are expected to have over ~60,000 samples.

For association studies, the immediate benefits of a larger panel include a more detailed catalog of genetic variants, which increases the chance of imputing a causal variant, and better imputation accuracy, which improves the power of downstream association analyses, especially for rare variants (Li et al. 2009b). On the other hand, the computational resources required to analyze such large panels using contemporary methods makes it cumbersome, if not infeasible (for example, imputing 1,000 GWA samples using the HRC reference panel requires ~2 years on a single CPU, or ~1 week on a 100-core cluster, using minimac2). Furthermore, panels such as the HRC and TOPMed have data restriction issues which inhibit users from gaining direct access to individual genotypes. This restriction prevents research groups from directly downloading these reference panels or merging them with their in-house private reference panels for improved imputation accuracy, thus diminishing their use as a next generation imputation resource. This dissertation illustrates the creation of a reference panel (Chapter 2) as well as a web server based method that allows imputation while still abiding by data sharing rules (Chapter 3).

### 1.1.3 Improving Imputation Quality - Trends and Strategies

Presently, the quality of genotype imputation depends more on the nature of the reference panel and the study design, and to a much lesser extent on the type of imputation algorithm. For example, when imputing against the HRC reference panel, the maximum absolute difference in imputation $r^2$ was 0.02 across all minor allele frequency (MAF) bins between minimac2, Impute2 and Beagle4.1 (Das et al. 2016, Table 1). This is because, irrespective of the latent algorithm, the underlying

mechanism behind any imputation tool is to identify stretches of shared haplotype segments, and most contemporary tools are almost equally efficient in doing this (Browning et al. 2016). While disparities in accuracy across different tools is more prominent with smaller reference panels, they become negligible when imputing against modern panels with large number of samples (Das et al. 2016, Table 1).

Unlike classical statistical problems, increasing the sample size is not the only way to improve imputation accuracy: taking into account the genetic nature of the data can yield additional benefits. The most intuitive strategies are increasing the size of the reference panel and increasing the coverage of the genotyping array. Both these approaches improve the chances of finding a possible match (either by increasing the number of templates to match against or increasing the number of SNPs to match with) which consequently improves imputation quality. For example, the Illumina® Human Omni5 Exome array is a better choice for imputation than the Human Omni2.5 Exome array, which in turn is better than the CoreExome array. Similarly, for European samples, using the HRC panel is a better a choice than either of the 1000G panels (see Figure 2.2, p.21). However, the HRC panel, being more enriched than the 1000G panel with haplotypes from Europeans, decreases imputation accuracy in non-European samples (see Figure H.3, p.103 in Appendix E). This counter example of the notion that increasing the reference panel always increases the imputation accuracy highlights the importance of genetic similarity between the study samples and the reference panel.

**Current Strategies.** The most common strategy for imputation is using a single publicly available dataset (such as the 1000 Genomes or the HRC) and it works quite well for studies with samples of European ancestry. However studies on non-European populations or relatively homogenous European sub-populations often benefit from using custom reference panels of samples with greater genetic similarity. For instance, in a study on samples in Sardinia, an island in Italy with a genetically isolated population, a custom reference of Sardinian samples provided better imputation accuracy than the 1000G and other large European panels (Pistis et al. 2014). This has also been replicated in other studies on homogenous European sub-populations (Deelen et al.

2014) as well as non-European populations such as African-Americans (Duan et al. 2013).

Additionally, disease specific studies often benefit from a hybrid approach that combines a custom reference panel with an existing public reference panel. This approach is particularly beneficial for disease studies as it significantly increases the accuracy of imputing a causal variant (since the hybrid panel is enriched with more copies of rare alleles). An example highlighting this feature: a recent study on prostate cancer was eventually able to impute the rare *HOXB13* G84E variant after using a hybrid imputation approach that combined the 1000 Genomes data with an enriched set of cases carrying the mutation (Hoffmann et al. 2015), even though previous studies had suggested that it might not be possible to impute this variant (Saunders et al. 2014). The hybrid method is also useful in studies with homogenous sub-populations given it enriches the panel with genetically similar samples while still retaining the benefits of large public panels.

**Multiple Reference Panels.** The hybrid approach requires studies to merge multiple reference panels and the streamlining of this process is still an active field of research. One proposed approach has been to combine the reference panels by first treating them as reference panels for each other and then "cross imputing" the missing variants into each reference (Howie et al. 2009). Although this approach enables one to use all variants found in any panel, the performance of this method has not been fully evaluated. One study found this approach to be neither helpful nor harmful for large, population-specific panels (Huang et al. 2015, Figure 1b). Some studies have repeated the imputation process for each of multiple different reference panels. For example, a study on sick sinus syndrome repeatedly imputed against HapMap, 1000G Pilot, and a set of genotyped samples to reveal a missense variant in the *MYH6* gene (c.2161C>T) associated with high risk. The ideal solution with multiple reference panels would be to call the variants jointly from their respective sequence alignment files for all the samples. However, variant calling is highly computationally intensive job and is not a feasible or practical solution for merging reference large panels. We address the issue of multiple reference panels in Chapter 5.

## 1.2 Outline of Dissertation

The second chapter of this dissertation describes our efforts to create the **Haplotype Reference Consortium** reference panel by combining information across 20 genetic studies as well as the reasoning for why these efforts were beneficial. We provide a brief summary of the methods used to combine the genetic data starting from the original alignment files, with details on the quality control and filtering methods, procedure for jointly calling the genotypes, and approaches for phasing the reference panel once genotypes have been called. We demonstrate improved imputation accuracy for European populations from the HRC panel (compared to 1000G panels) as well as the added benefit of strengthening association signals from known disease risk loci.

The third chapter describes an imputation tool for significantly faster analysis **(minimac3)** and a web platform service that considerably improves user experience **(Michigan Imputation Server)**. The web server also helps comply with data restriction issues. We explain the idea of state space reduction - analyzing only unique haplotypes in small genomic segments. We provide formulations for implementing an HMM with the reduced state space; proofs have been omitted from the main text and are detailed in Appendix D. We compare the imputation accuracy, total run-time, and physical memory requirement of minimac3 to other contemporary tools, including minimac2, Impute2, and Beagle4.1. We assess how the computational burden and memory requirement of minimac3 scales with larger panels. A brief description of the features implemented on the imputation server is also provided.

The fourth chapter further extends the idea of state space reduction to build even faster imputation tools that can scale to future reference panels **(minimac4)**. We describe a more aggressive version of state space reduction by collapsing haplotypes based on genotyped markers only. The necessary formulas are described in the methods, and the detailed proofs are provided in Appendix E. We also describe other software engineering techniques that were applied to further decrease the computational burden in minimac4. We compare the performance throughput of minimac4 to its previous version and to Impute2 and Beagle4.1. We also give predicted run-times for imputing

against some common reference panels. Finally, we discuss imputation accuracy in non-European studies and optimum reference panels for such studies.

The fifth chapter addresses issues related to merging reference panels. We introduce the idea of meta-imputation, in which study samples imputed against different reference panels are combined to obtain a consensus imputed result **(MetaMinimac)**. We give a brief summary of the intuition and formulas used in MetaMinimac. We demonstrate that meta-imputation performs equally well as joint imputation (in which the reference panels have been merged before imputation) in all the scenarios that we examined, including studies of samples of European, non-European, mixed, and admixed ancestry. We also demonstrate the benefit of meta-imputation in reducing computation time for imputation against large reference panels.

Minimac3/4 are subsequent versions in the MaCH/minimac family of imputation tools. Minimac4 is a more generalized version of the algorithms implemented in minimac3. Although minimac4 requires lesser time to impute each sample, it has an overhead cost and thus might not be beneficial for small sample sizes. Minimac3 might yield lesser overall computation time for studies with less than 100 individuals. Both these tools yield the same imputation accuracy and are as accurate as other existing tools. When a single reference panel is available (such as HRC), minimac3/4 can be used to impute into GWA samples. When multiple reference panels are available, investigators can impute their study samples against each reference panel and then combine the results together using metaMinimac. For studies of non-European samples and disease specific studies, the later approach is more beneficial.

The last chapter summarizes what has been accomplished in this dissertation.

# CHAPTER 2

## Haplotype Reference Consortium - A Reference Panel of 64,976 Haplotypes for Genotype Imputation[1]

### 2.1 Introduction

Over the last decade, large scale international collaborative efforts have created successively larger and more ethnically diverse genetic variation resources. For example, in 2007 the International HapMap Project produced a haplotype reference panel of 420 haplotypes at ~3.1M SNPs in 3 continental populations (Frazer et al. 2007). More recently, the 1000 Genomes Project has produced a series of datasets built using low-coverage WGS, culminating in 2015 in a reference panel of 5,008 haplotypes at over ~88M variants from 26 world-wide populations, namely the 1000G Phase 3 (1000G 2015). In addition, several other projects have collected low-coverage WGS data in large numbers of samples that could potentially also be used to build haplotype reference panels (Francalacci et al. 2013; GoNL 2014; Huang et al. 2015). A major use of these resources has been to facilitate imputation of unobserved genotypes into GWA study samples that have been assayed using relatively sparse genome-wide microarray chips. As the reference panels have increased in number of haplotypes, SNPs and populations, genotype imputation accuracy has increased, allowing researchers to impute and test SNPs for association at ever lower minor allele frequencies.

### 2.2 Methods

The HRC was formed to bring together as many WGS datasets as possible to build a much

[1]McCarthy, S., **S. Das**, W. Kretzschmar, O. Delaneau, A. R. Wood, et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature Genetics* 48.10, pp. 1279–1283

larger combined haplotype reference panel. By doing so, our aim is to provide a single centralized resource for human genetics researchers to carry out genotype imputation. Here we describe the first HRC reference panel that combines datasets from 20 different studies. Details on the studies are provided in Table C.2, p.81 in Appendix C. Most of these studies have low-coverage WGS data (4-8X coverage) and are known to consist of samples with predominantly European ancestry. However, the 1000G Phase 3 cohort, which has diverse ancestry, is also included. This reference panel consists of $64,976$ haplotypes at $39,235,157$ SNPs that have evidence of having a minor allele count (MAC) greater or equal to 5.

### 2.2.1 Quality Control

We took the following approach to create the reference panel. We combined existing sets of genotype calls from each study to determine a 'union' set of $95,855,206$ SNP markers with MAC $\geq$ 2 (MAC2 list). After initial tests, we decided for this first version of the HRC panel not to include small insertions and deletions (indels), since these were very inconsistently called across projects. We then used a standard tool to calculate the genotype likelihoods consistently for each sample at each marker from the original study binary alignment map (BAM) files (Li et al. 2009a) and make a baseline set of non-linkage disequilibrium (LD) based genotype calls. We next applied several filters to remove poor quality markers (Figure C.1, p.83 in Appendix C). We restricted filters to markers with MAC $\geq$ 5 (MAC5 list), corresponding to a minimum MAF of 0.0077%, then added back markers that are present on several commonly used SNP microarray chips in GWA studies. After filtering, this marker list consisting of $44,187,567$ markers exhibited improved quality compared to MAC $\geq$ 5 marker list when assessed by measuring a per sample transition-to-transversion (Ts/Tv) ratio (Figure 2.1). We also detected and removed 301 duplicate samples across the whole dataset (Table C.3, p.82 in Appendix C).

### 2.2.2 Phasing

Calling genotypes and phasing using low-coverage WGS data has been a computational chal-

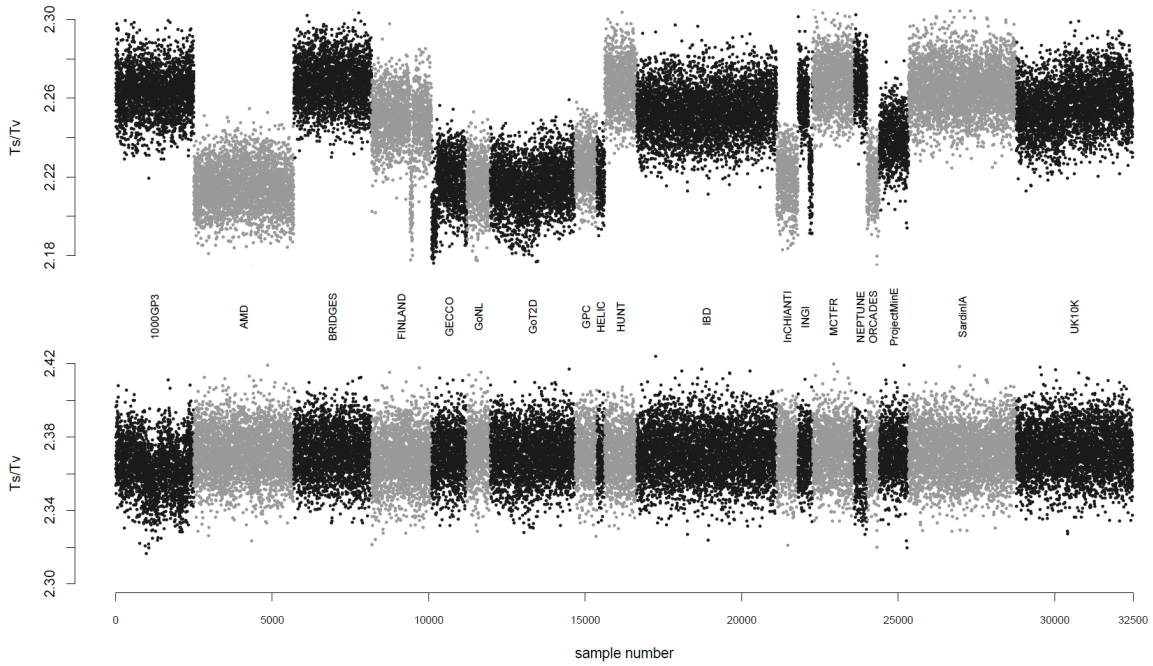


Figure 2.1: Effect of variant filtering on Ts/Tv ratio in the HRC panel. The top figure shows the per-sample Ts/Tv ratio for chromosome 20 on the full MAC5 marker list. The bottom figure shows the same after the marker filtering steps

lenging step for many of the 20 studies providing data. To reduce computation, we carried out this step on all $32,611$ samples together and leveraged the original separately called haplotypes from each study to help reduce the search space of the calling algorithm. We then applied a further refinement step by re-phasing the called genotypes using the SHAPEIT method (Delaneau et al. 2012), based on experience from the UK10K project, which found this re-phasing approach produced substantially improved imputation accuracy when using the haplotypes (Huang et al. 2015). After final genotype calling, we removed a further 123 samples and filtered out 4,952,410 markers whose MAC after refinement and sample removal was below 5, resulting in a final set of $39,235,157$ markers and $32,488$ samples.

## 2.3 Results

We evaluated accuracy of our genotype calls by measuring their genotype discordance compared to Illumina® Omni2.5M chip genotypes available on the 1000 Genomes samples. We found

that both our marker filtering strategy and the increased sample size of HRC led to improved accuracy (Table C.1, p.80 in Appendix C). For example, we obtained a non-reference allele discordance of 39% on the full HRC dataset with marker filtering, compared to 67% on the subset of 1000G Phase 3 samples.

### 2.3.1 Improvement in Imputation Accuracy

We next carried out experiments to assess and illustrate the downstream imputation performance compared to previous haplotype reference panels. To mimic a typical imputation analysis, we created a pseudo-GWA study dataset using high-coverage Complete Genomics (CG) WGS genotypes on 10 European samples. We extracted the CG SNP genotypes at all the markers included on an Illumina® 1M SNP array (Human1M-Duo BeadChip). These were used to impute the remaining genotypes which were then compared to the held out genotypes. We calculated aggregate[2] $r^2$ by stratifying results by alternate allele frequency (AF) of the imputed markers. Figure 2.2 shows the results of imputation using two different tools: minimac3 and PBWT (Das et al. 2016; Durbin 2014). We see that the HRC reference panel leads to a large increase in imputation performance compared to the 1000G Phase 3 ($r^2 = 0.70$ vs $r^2 = 0.50$ for minimac3 and $r^2 = 0.64$ vs $r^2 = 0.25$ for PBWT at alternate AF = 0.1%). HRC imputation at 0.1% frequency provides similar accuracy to 1000G Phase 3 imputation at 1% frequency.

### 2.3.2 Higher Resolution for Fine Mapping

To further illustrate the benefits of using the HRC resource, we imputed a GWA studies study of 1,210 samples from the InCHIANTI study (Ferrucci et al. 2000), including 534 that did not contribute to the HRC reference panel because they were not sequenced. Imputing using the HRC

[2]An alternative to average $r^2$ when relatively small number of samples (~10) are used for imputation accuracy experiments. Instead of calculating Pearson $r^2$ at each variant with only ~10 samples and then averaging across a fixed alternate allele frequency bin, we combine all variants in the fixed bin and calculate a single Pearson $r^2$ across all genotypes in this bin. Aggregate $r^2$ has lesser squared error in estimating true imputation $r^2$ when there are small number of study samples.

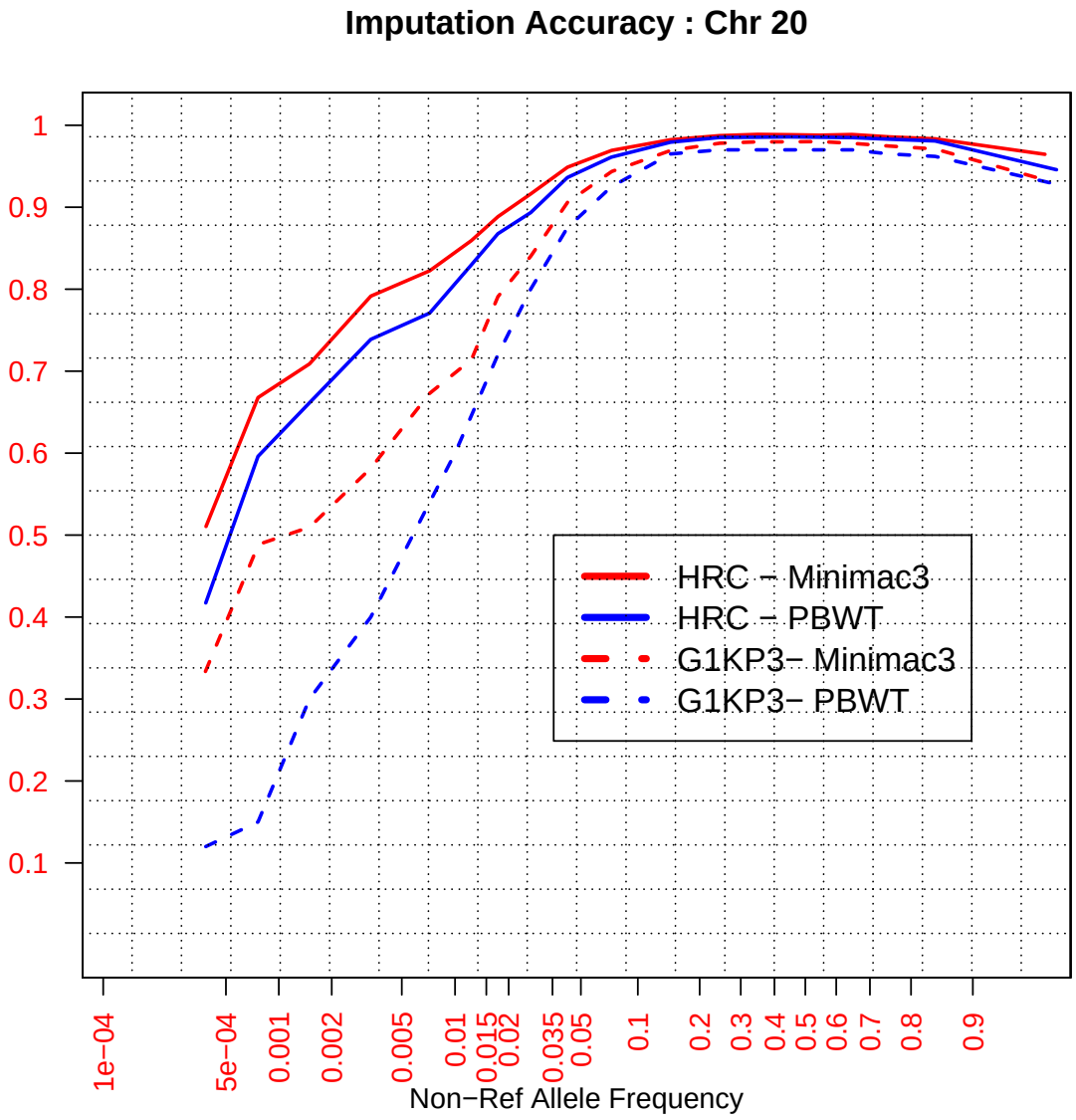


Figure 2.2: Improved imputation accuracy from the HRC panel. We compare accuracy between HRC and 1000G Phase 3 (abbreviated as G1KP3) using minimac3 and PBWT. The X-axis shows the non-reference allele frequency on a log scale. The Y-axis shows accuracy measured by aggregate $r^2$ when imputing SNP genotypes into 10 CEU samples

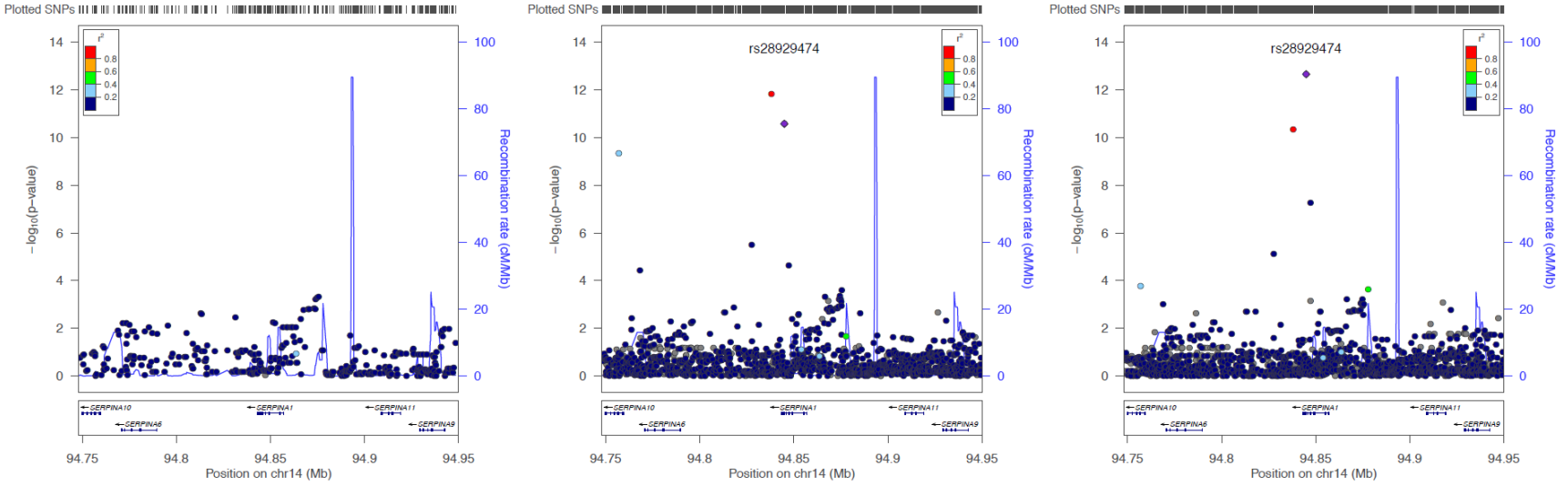


Figure 2.3: Improved association signal from the HRC panel. Association test statistics on the –log10 p-value scale (y-axis) are plotted for each SNP position (x-axis). Three different imputation panels were used: HapMap2 (left), 1000G Phase 3 (middle), HRC release 1 (right). The SNP rs28929474 is shown as a purple. HRC imputation gives a clear refinement of the signal at the rare causal SNP rs28929474 (MAF=0.5%)

panel resulted in ∼15M SNPs passing an imputation quality threshold of $r^2$ >0.5 compared to ∼13M variants when imputing using 1000 Genomes Phase 3, an increase of over 2 million variants. The majority of these additional SNPs occur at the lower frequency range. We next tested the HRC imputed genotypes for association with 93 circulating blood marker phenotypes, including many of relevance to human health such as lipids, vitamins, ions, inflammatory markers and adipocytes. This analysis highlighted potential novel associations and that it is possible for HRC imputation to refine signals of association (detailed results omitted). For example, Figure 2.3 shows the association results of HapMap2, 1000G Phase 3 and HRC based imputation for the $\alpha$1-antitrypsin phenotype at the *SERPINA1* locus. HRC imputation gives a clear refinement of the signal at the rare causal SNP rs28929474 (MAF=0.5%), known to predispose to the $\alpha$1-antitrypsin deficiency lung condition emphysema (Bathurst et al. 1984; Ferrarotti et al. 2012).

## 2.4 Discussion

This first release of the HRC is the largest human genetic variation resource to date and has been created via an unprecedented collaboration of data sharing across many groups. We envisage continuing to expand the HRC and are currently planning a second HRC release differing from the first release in two ways. Firstly, we aim to substantially increase the ethnic diversity of the panel, by including data from sequencing studies in world-wide sample sets such as the CONVERGE study (CONVERGE 2015), AGVP (Gurdasani et al. 2014) and HGDP (Rosenberg et al. 2002). Secondly, we aim to include short insertions and deletions in addition to SNP variants.

# CHAPTER 3

## Minimac3 : Next-generation Genotype Imputation Service and Methods[1]

### 3.1 Introduction

Genotype imputation means estimation of genotypes or genotypes probabilities at markers that have not been directly genotyped in a typical genetic study. After study samples are genotyped on a commercial DNA array, typically directly assaying $300,000 - 1,000,000$ SNPs, imputation works by finding haplotype segments that are shared between study individuals and a reference panel of sequenced genomes, such as those from the 1000G Project (1000G et al. 2012; 1000G 2015), or recent population sequencing studies (Francalacci et al. 2013; GoNL 2014; Gudbjartsson et al. 2015). Imputation accurately assigns genotypes at untyped markers, improving genome coverage, facilitating comparison and combination of studies that use different marker panels, guiding fine-mapping, and increasing power to detect genetic association (Li et al. 2009b; Marchini et al. 2010).

The accuracy of imputation increases with the number of haplotypes in the reference panel of sequenced genomes (Fuchsberger et al. 2014; Howie et al. 2012), particularly for rare (MAF $\leq 0.5\%$) and low-frequency ($0.5\% <$ MAF $< 5\%$) variants. These rare and low-frequency variants include most loss-of-function alleles (MacArthur et al. 2012) and other high-impact variants, identification of which enable genotype-based callback and focused studies of natural knockout alleles (MIGCI et al. 2014; Sulem et al. 2015). Large reference panels, such as the one developed by the HRC are expected to extend accurate imputation to variants with frequencies of $0.1 - 0.5\%$

[1] **Das, S.**, L. Forer, S. Schonherr, C. Sidore, A. Locke, et al.(2016).“Next-generation genotype imputation service and methods”. In: *Nature Genetics* 48.10, pp. 1284–1287

or less, and already include 1,000s of putative loss-of-function and protein-truncation alleles. The HRC panel combines sequence data across >32,000 individuals from 20 medical sequencing studies, and is cumbersome to access directly because of participant privacy protections in individual studies and the large volumes of data involved. Using current well established approaches (Fuchsberger et al. 2014), we estimate imputing $1,000$ GWA study samples using the HRC reference set would require >4 years on a single CPU.

In this chapter, we present new algorithms for genotype imputation that increase computational efficiency with no loss of accuracy by leveraging local similarities between sequenced haplotypes. We also present a new model for imputation, based on a web-service that greatly simplifies analysis, eliminates the need for cumbersome data access agreements, and thus allows researchers to devote their time to more interesting tasks. The approach described here represents the current most efficient strategy for genotype imputation and has been used to process >7.5 million genotyped samples through our imputation server. Our implementation is freely available, enabling others to build on our advances and ideas.

## 3.2 Methods

Our improved imputation algorithm is based on a 'state-space reduction' of the HMM that is typically used to describe haplotype sharing (Li et al. 2010); it exploits similarities among haplotypes in small genomic segments to reduce the effective number of states over which the HMM iterates (Figure 3.1). Our model divides the genome into consecutive blocks and iterates only over the unique haplotypes in each genomic block. It then uses a reversible mapping function that can reconstruct exactly the same state-space used by minimac and Impute2 (Howie et al. 2012). An important feature of our derivation is that yields exactly the same imputation results as more cumbersome analyses in the original state-space. We implemented this method in a `c++` package called **minimac3**.

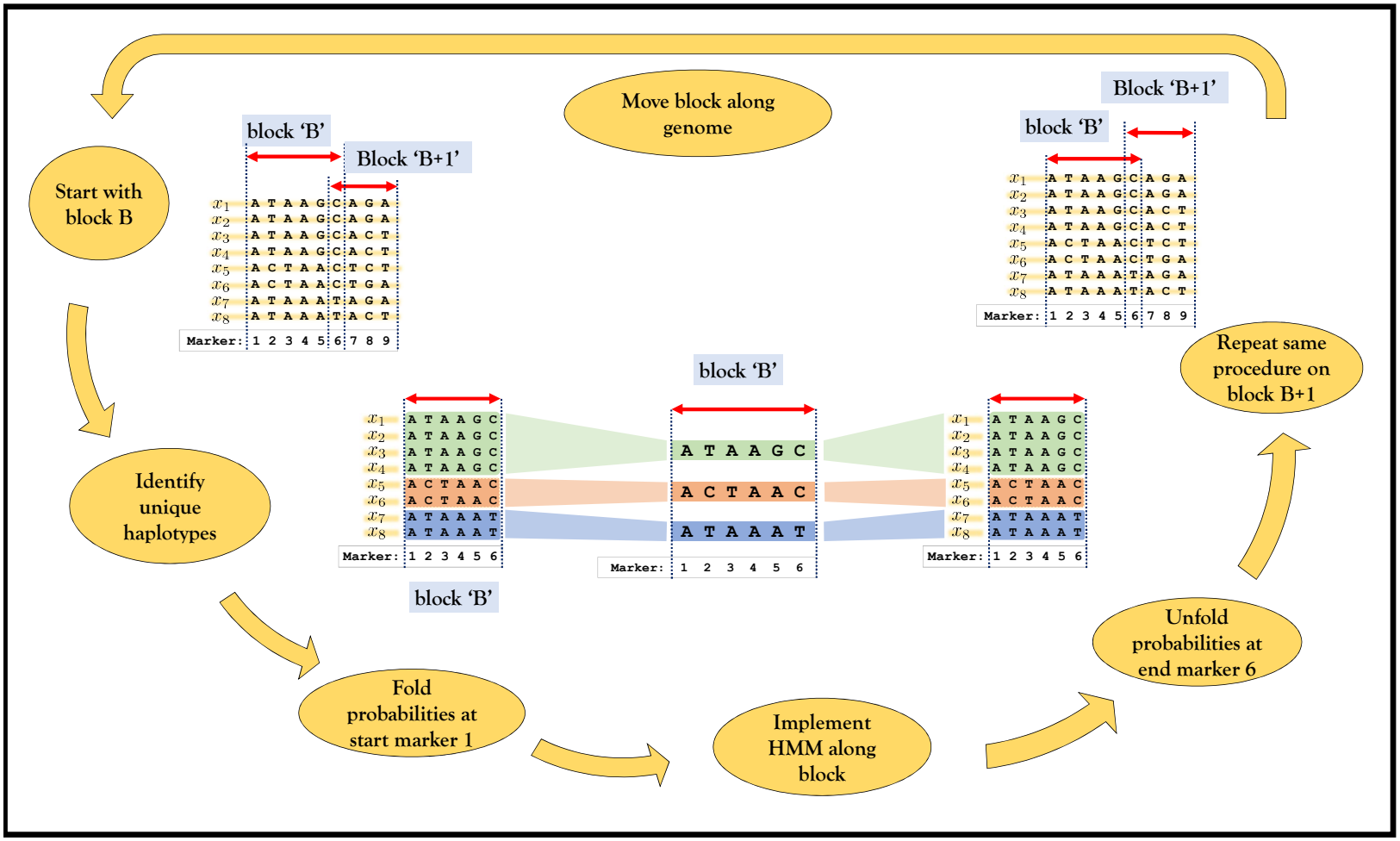


Figure 3.1: Overview of state space reduction in minimac3. We consider a chromosome region with $M = 9$ markers and $H = 8$ haplotypes $x_1$, $x_2$, ..., $x_8$. (A) We break the region into consecutive genomic segments (blocks) and start by analyzing Block $B$ from marker 1 to 6. (B) In Block $B$ we identify $U = 3$ unique haplotypes $Y_1$, $Y_2$ and $Y_3$ (colored in green, red, and blue). Given the left probabilities of the original state space at marker 1 (i.e. $L_1(x_1), \ldots, L_1(x_8)$), we fold them to get left probabilities of the reduced state space at marker 1 (i.e. $\boldsymbol{\mathcal{L}}_1(y_1)$,$\boldsymbol{\mathcal{L}}_1(y_2)$ and $\boldsymbol{\mathcal{L}}_1(y_3)$). (C) We implement HMM on the reduced state space ($Y_1$, $Y_2$ and $Y_3$) from marker 1 to marker 6 to get $\boldsymbol{\mathcal{L}}_6(y_1)$,$\boldsymbol{\mathcal{L}}_6(y_2)$ and $\boldsymbol{\mathcal{L}}_6(y_3)$. (D) We next unfold the left probabilities of the reduced state space at marker 6 to obtain the left probabilities of the original state space (i.e. $L_6(x_1), \ldots, L_6(x_8)$). (E). We repeat the same procedure on the next block, starting with $L_6(x_1), \ldots, L_6(x_8)$ to finally obtain $L_9(x_1), \ldots, L_9(x_8)$.

### 3.2.1 State Space Reduction

In this section, we describe the state space reduction that uses the similarity between haplotypes in small genomic segments to reduce computational complexity. We recommend first reading a description of the original MaCH/minimac algorithm (Howie et al. 2012). For a brief review see Section A.2, p.74 of Appendix A). In the subsections below, we first describe the modified forward equations, backward equations and final imputation formula. Detailed mathematical proofs for these equations are provided in Appendix D.

Consider a reference with $H$ haplotypes and a genomic segment bounded by markers $P$ and $Q$. Let $U \leq H$ be the number of distinct haplotypes in the block. Label the original haplotypes as $X_1, X_2, \ldots, X_H$ (the original space) and the distinct unique haplotypes as $Y_1, Y_2, \ldots, Y_U$ (the reduced space). For example, in Figure 3.1, the block $B$ bounded by markers $P = 1$ to $Q = 6$ has $U = 3$ distinct haplotypes and $H = 8$ actual haplotypes.

**Forward Equations** Let $L_k(.)$ (and $\boldsymbol{\mathcal{L}}_k(.)$) denote the left probabilities for the original states (and reduced states) at marker '$k$' (for $P \leq k \leq Q$) (Howie et al. 2012; Li et al. 2010). Below, we have three sets of equations: the first equation (3.1) explains how we transform the left probabilities from the original space to the reduced space ($L_P \rightarrow \boldsymbol{\mathcal{L}}_P$), the second equation (3.2) gives the modified formulation for the Baum-Welch forward equation for the reduced state space ($\boldsymbol{\mathcal{L}}_P \rightarrow \boldsymbol{\mathcal{L}}_Q$), and the third equation (3.3) transforms the left probabilities from the reduced space to the original space ($\boldsymbol{\mathcal{L}}_Q \rightarrow L_Q$).

$$\boldsymbol{\mathcal{L}}_P(Y_i) = \sum_{j=1}^{H} I(X_j = Y_i) \times L_P(X_j) \tag{3.1}$$

$$\boldsymbol{\mathcal{L}}_k\left(Y_i\right) = \left[\left[1 - \lambda_{k-1}\right] \boldsymbol{\mathcal{L}}_{k-1}\left(Y_i\right) + \frac{N_i \lambda_{k-1}}{H} \sum_{j=1}^{U} \boldsymbol{\mathcal{L}}_{k-1}\left(Y_j\right)\right] \times P\left(S_k | Y_i\right) \tag{3.2}$$

$$L_Q(X_j) = \boldsymbol{\mathcal{L}}_Q^R(Y_i) \times \left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{L}}_Q^{NR}(Y_i) \left[\frac{L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right] \tag{3.3}$$

In the above equations:

- $\lambda_k$ denotes the template switch probability between markers $k$ and $k+1$ (analogous to a "recombination fraction")
- $S_k$ the genotype in the study sample and $P\left(S_k|Y_i\right)$ denotes the genotype emission probabilities (is equal to 1 if study sample is missing at the marker) (Li et al. 2010)
- $N_i$ the number of haplotypes from the original state space that were collapsed to form $Y_i$ (i.e. $\sum_{i=1}^{U} N_i = H$)
- $\boldsymbol{\mathcal{L}}_k^{NR}(.)$ and $\boldsymbol{\mathcal{L}}_k^{R}(.)$, respectively, are components of the left probabilities assuming no recombination event or at least one recombination event between markers 1 and $k$. They are defined as follows (for each $k$):

$$\boldsymbol{\mathcal{L}}_k^{NR}(Y_i) = \boldsymbol{\mathcal{L}}_P(Y_i) \prod_{i=P}^{k-1} \left[(1-\lambda_i) P\left(S_{i+1}|Y_i\right)\right] \tag{3.4}$$

$$\boldsymbol{\mathcal{L}}_k^{R}(Y_i) = \boldsymbol{\mathcal{L}}_k(Y_i) - \boldsymbol{\mathcal{L}}_k^{NR}(Y_i) \tag{3.5}$$

**Backward Equations** Similarly, if $R_k(.)$ (and $\boldsymbol{\mathcal{R}}_k(.)$) denote the right probabilities for the original states (and reduced states) at any marker '$k$' (for $P \leq k \leq Q$), the following equations respectively transform the right probabilities $R_Q \rightarrow \boldsymbol{\mathcal{R}}_Q$ (3.6), gives the modified formulation for the Baum-Welch backward equations $\boldsymbol{\mathcal{R}}_Q \rightarrow \boldsymbol{\mathcal{R}}_P$ (3.7), and then transforms back the right probabilities $\boldsymbol{\mathcal{R}}_P \rightarrow R_P$ (3.8).

$$\boldsymbol{\mathcal{R}}_Q(Y_i) = \sum_{j=1}^{H} I(X_j = Y_i) \times R_Q(X_j) \tag{3.6}$$

$$\boldsymbol{\mathcal{R}}_k\left(Y_i\right) = \frac{N_i \lambda_k}{H} \left[\sum_{j=1}^{U} \boldsymbol{\mathcal{R}}_{k+1}\left(Y_j\right) P\left(S_{k+1}|Y_j\right)\right] + \left[1-\lambda_k\right] \boldsymbol{\mathcal{R}}_{k+1}\left(Y_i\right) P\left(S_{k+1}|Y_i\right) \tag{3.7}$$

$$R_P(X_j) = \boldsymbol{\mathcal{R}}_P^{R}(Y_i) \times \left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{R}}_P^{NR}(Y_i) \left[\frac{R_Q(X_j)}{\boldsymbol{\mathcal{R}}_Q(Y_i)}\right] \tag{3.8}$$

In the above equations, $\boldsymbol{\mathcal{R}}_k^{NR}(.)$ and $\boldsymbol{\mathcal{R}}_k^{R}(.)$, respectively, are components of the right probabilities assuming no recombination event or at least one recombination event between markers $k$ and

$Q$. They are defined as follows (for each $k$):

$$\boldsymbol{\mathcal{R}}_k^{NR}(Y_i) = \boldsymbol{\mathcal{R}}_Q(Y_i) \prod_{i=k}^{Q-1} \left[(1-\lambda_i) P\left(S_{i+1}|Y_i\right)\right] \tag{3.9}$$

$$\boldsymbol{\mathcal{R}}_k^{R}(Y_i) = \boldsymbol{\mathcal{R}}_k(Y_i) - \boldsymbol{\mathcal{R}}_k^{NR}(Y_i) \tag{3.10}$$

**Final Imputation Formula** The final formula that gives the posterior probability of each reduced state in the given genomic block is given in equation (3.11). The mathematical details behind the proof have been explained in Appendix D.

$$\begin{aligned} P\left(Y_i\,|\text{GWAS}\right) \;=\; & \left[\sum_{j=1}^{H} I(X_j = Y_i) L_P\left(X_j\right) R_Q\left(X_j\right)\right] \times \left[\frac{\boldsymbol{\mathcal{L}}_K\left(Y_i\right)}{\boldsymbol{\mathcal{L}}_P\left(Y_i\right)} \times \frac{\boldsymbol{\mathcal{R}}_K\left(Y_i\right)}{\boldsymbol{\mathcal{R}}_Q\left(Y_i\right)}\right] \\ & + \frac{1}{N_i}\left[\boldsymbol{\mathcal{L}}_K\left(Y_i\right)\boldsymbol{\mathcal{R}}_K\left(Y_i\right) - \boldsymbol{\mathcal{L}}_K^{NR}\left(Y_i\right)\boldsymbol{\mathcal{R}}_K^{NR}\left(Y_i\right)\right] \end{aligned} \tag{3.11}$$

### 3.2.2 Computational Complexity and Optimal Allocation

Methods that perform phasing and imputation simultaneously (for example, MaCH (Li et al. 2010) and Impute (Marchini et al. 2007)) have a computational cost proportional to number of study samples ($N$), number of genotyped markers in the reference panel ($M$), and square of number of reference haplotypes ($H^2$) or $O(NMH^2)$. However, in the context of pre-phasing, as in minimac and Impute2 (Howie et al. 2012) this becomes $O(NMH)$.

In this study, we break up the whole chromosome into $K$ consecutive segments. If $U_i$ and $M_i$ denote the number of unique haplotypes and markers in each segment, then complexity is $O\left(N \times \sum_{i=1}^{K} U_i M_i\right) + O(NKH)$. The second term accounts for the complexity of transitions between blocks, which occur in the original state space. Thus, although very short segments could reduce the number of unique haplotypes per segment ($U_i$) and complexity measured by the first term, these would also increase the total number of segments ($K$) and complexity measured by the second term. An optimal allocation of genomic regions must balance these two goals.

We implement a recursive dynamic programming algorithm to find the optimal allocation of the

genomic segments, since brute is not feasible ($\sim 2^{M-1}$ alternatives). We assume that the optimal complexity of imputation until marker $i < M$ is denoted by $C(i)$ and calculate $C(M)$ recursively as in equation (3.12):

$$C(M) \quad = \quad \min_{i=1,2,\ldots,M-1} \{C(i) + [U(i,M)(M-i+1)] + 2H\} \tag{3.12}$$

In this expression, $C(i)$ is the optimal cost for imputation from marker 1 to marker $i$, $U(i,M)(M-i+1)$ is the cost of imputing the genomic segment from marker $i$ to marker $M$, and $2H$ is the transitioning between segments. This expression would require at most $M^2$ comparisons and this number can be further reduced since we can rule out very long segments.

### 3.2.3 Comparison of Minimac2, Minimac3, Impute2, and Beagle4.1

We evaluated the performance of Minimac3 in comparison to the three most commonly used imputation tools – minimac2 (Fuchsberger et al. 2014), Impute2 (Howie et al. 2012), and Beagle v4.1 (Browning et al. 2016). We combined chromosome 20 data across multiple whole genome sequencing studies to generate large reference panels. We compared the six reference panels described in Table 3.1, p.31. To get a consensus set of variants, we imputed into only those variants which occurred in all contributing studies, yielding a total of $227,925$ variants (with a minimum minor allele frequency of $5 \times 10^{-5}$). To mimic a GWA study, we selected 25 unrelated samples each from AMD, SARD, BRIDGES, and TWINS and masked all variants except those typed on the Illumina® Duo 1M Chip (resulting in ∼20K genotyped variants for chromosome 20). To evaluate imputation accuracy, we estimated the squared Pearson correlation coefficient between the imputed genotype probabilities and hard genotype calls from sequenced data at the masked variants. For each of the four imputation methods and six reference panels combination, we recorded the aggregate[2] $r^2$, total computational time, and physical memory required to impute 100 GWA study individuals.

[2]See footnote on page 20 for definition of aggregate $r^2$

| | |
|---|---|
| G1KP1 | 1,092 individuals from 1000G Phase 1 |
| AMD | 2,074 individuals sequenced for study of age-related macular degeneration |
| G1KP3 | 2,504 individuals from the 1000G Phase 3 |
| SARD | 3,489 individuals from Sardinia |
| COMB | 9,341 individuals from AMD; SARD; 2,464 individuals sequenced for a study of bipolar disorder (BRIDGES); and 1,314 individuals sequenced for Minnesota Twins study (TWINS) |
| MEGA | 11,845 individuals from COMB and 2,504 from G1KP3 |

Table 3.1: Description of reference panels used for imputation experiments.

## 3.3 Results

### 3.3.1 Faster Imputation

The results from our imputation experiment is shown in Table 3.3, p.33. Imputation using Impute2 required the most running time (over 2 weeks). For minimac2 and Impute2, memory and run time increased linearly with panel size. Increasing panel size by 10-fold from $1,091$ to $11,455$, memory required increased 10-fold from 0.3 to 3.5 GB for minimac2 and from 0.9 to 8.3 GB for Impute2, while run time increased from 26 to 304 hours for minimac2 and 34 to 363 hours for Impute2. Beagle4.1 did better than linear as its run time increased from 5 to 40 hours (8-fold). Minimac3 consistently outperformed all alternatives: increasing panel size 10-fold from $1,091$ to $11,455$ samples increased memory requirements 3.5 times and run time 5-fold. For the largest panel (11,455 samples), minimac3 was twice as fast than Beagle4.1, 14 times faster than minimac2, and 17 times faster than Impute2, and reduced memory usage by 23%, 90%, and 96%, respectively (Table 3.3, p.33). In this comparison, all programs were run using a single thread, allowing for fair comparison. For a comparison of wall-clock running times, when some methods are run in a single thread and others in multiple threads, see Browning and Browning (Browning et al. 2016). We compared imputation quality across the four methods by calculating the squared correlation coefficient $r^2$ between imputed allele dosages and masked genotypes (Table 3.4, p.34). Minimac, minimac2, and minimac3 are based on the same mathematical model and hence gave identical

| **Reference panel sample size** | **Imputation accuracy** | | | **Probability an individual judged by imputation to carry the rare allele actually does carry that allele** | | |
|---|---|---|---|---|---|---|
| | Minor allele frequency bin | | | | | |
| | $10^{-4}$ - 0.01% | .01 - 1% | >1% | $10^{-4}$ - .01% | .01 - 1% | >1% |
| 1,000 | 0.34 | 0.56 | 0.91 | 0.48 | 0.57 | 0.89 |
| 2,000 | 0.39 | 0.59 | 0.93 | 0.52 | 0.60 | 0.90 |
| 5,000 | 0.46 | 0.64 | 0.94 | 0.57 | 0.64 | 0.92 |
| 10,000 | 0.51 | 0.68 | 0.95 | 0.60 | 0.67 | 0.93 |
| 20,000 | 0.60 | 0.74 | 0.96 | 0.67 | 0.75 | 0.94 |

Table 3.2: Imputation accuracy based on simulated haplotypes. The haplotypes were simulated under a coalescent model consistent with European haplotype diversity, and imputed into GWA study data.

results. Minimac3 slightly outperformed Beagle4.1 and moderately outperformed Impute2, particularly for the rare variant MAF bin $[10^{-4} - 0.5\%]$ where, with $3,489$ reference samples, Impute2 attained $r^2 = 53.3\%$ while Beagle4.1 attained $54.3\%$, and minimac3 attained $55.5\%$. All methods demonstrated improved imputation quality with increasing panel size, particularly for rare variants. For example, for minimac3, quality increased from $r^2 = 45\%$ to $r^2 = 77\%$ when panel size increased from $1,092$ to $11,455$.

### 3.3.2 Scaling with Large Panels

To evaluate the benefits of large panels, we used coalescent simulations with up to 20,000 reference panel individuals. We found substantial gains in imputation accuracy between imputed genotypes and the true simulated genotypes, as panel size increased (Table 3.2). For variants with (MAF $\leq 0.001\%$), average imputation $r^2$ increased from 0.41 to 0.79 when reference panels grew from $N = 1,000$ to $N = 20,000$ individuals; this represents a near 2-fold increase in effective sample size for association (which scales as $r^2$) (Pritchard et al. 2001). For example, with a reference panel of 20,000 individuals, our simulations estimate the probability individuals identified as a rare allele carrier through imputation actually carry the allele is ~84% (MAF $0.01 - 0.1\%$)

| Reference Panel | # Samples | minimac3 | minimac2 | Impute2 | Beagle 4.1 |
|---|---|---|---|---|---|
| Time (in CPU-hours) | | | | | |
| G1KP1 | 1,092 | **4** | 27 | 34 | 5 |
| AMD | 2,074 | **9** | 59 | 73.5 | 9 |
| G1KP3 | 2,504 | **6** | 61 | 78 | 9 |
| SARD | 3,489 | **7** | 85 | 108 | 11 |
| COMB | 9,341 | **17** | 236 | 288 | 31 |
| MEGA | 11,845 | **21** | 304 | 364 | 40 |
| Memory (in CPU-Gigabytes) | | | | | |
| G1KP1 | 1,092 | **0.09** | 0.34 | 0.91 | 0.51 |
| AMD | 2,074 | **0.14** | 0.62 | 1.58 | 0.39 |
| G1KP3 | 2,504 | **0.13** | 0.75 | 1.88 | 0.56 |
| SARD | 3,489 | **0.13** | 1.03 | 2.55 | 0.46 |
| COMB | 9,341 | **0.28** | 2.73 | 6.57 | 0.41 |
| MEGA | 11,845 | **0.33** | 4.61 | 8.28 | 0.43 |

Table 3.3: Computational requirement of minimac3 and other imputation tools. The tables summarizes the running time and memory requirements of minimac3, minimac2, Impute2, and Beagle4.1 for different reference panels to impute 100 whole genomes (running time interpolated from analysis on chromosome 20). All four tools were run on 5 Mb chunks with 1 Mb overlap (13 chunks in serial or chromosome 20 yielding a total of 227,925 variants). Minimac3, minimac2, Impute2 were run with pre-calculated estimates. Minimac3 and Beagle4.1 was run with their respective input file formats (m3vcf)[3] and bref[4] respectively) while minimac2 and Impute2 were run on VCF and Oxford format files respectively. The best results are highlighted in bold face.

| Reference Panel | # Samples | minimac3 | minimac2 | Impute2 | Beagle 4.1 |
|---|---|---|---|---|---|
| Imputation Accuracy (Mean $r^2$) MAF [$10^{-4}$-0.5%] | | | | | |
| G1KP1 | 1,092 | **0.45** | **0.45** | 0.43 | 0.42 |
| AMD | 2,074 | **0.54** | **0.54** | 0.51 | 0.52 |
| G1KP3 | 2,504 | **0.52** | **0.52** | 0.49 | **0.52** |
| SARD | 3,489 | **0.55** | **0.55** | 0.53 | 0.54 |
| COMB | 9,341 | **0.76** | **0.76** | 0.74 | **0.76** |
| MEGA | 11,845 | **0.76** | **0.76** | 0.74 | **0.76** |
| Imputation Accuracy (Mean $r^2$) AF [0.5-5%] | | | | | |
| G1KP1 | 1,092 | **0.77** | **0.77** | 0.76 | 0.73 |
| AMD | 2,074 | **0.82** | **0.82** | 0.80 | 0.80 |
| G1KP3 | 2,504 | **0.79** | **0.79** | 0.78 | **0.79** |
| SARD | 3,489 | 0.79 | 0.79 | 0.78 | **0.80** |
| COMB | 9,341 | **0.89** | **0.89** | 0.88 | **0.89** |
| MEGA | 11,845 | **0.89** | **0.89** | 0.88 | **0.89** |
| Imputation Accuracy (Mean $r^2$) MAF [5-50%] | | | | | |
| G1KP1 | 1,092 | **0.96** | **0.96** | 0.95 | 0.95 |
| AMD | 2,074 | **0.96** | **0.96** | **0.96** | **0.96** |
| G1KP3 | 2,504 | **0.96** | **0.96** | **0.96** | **0.96** |
| SARD | 3,489 | **0.96** | **0.96** | **0.96** | **0.96** |
| COMB | 9,341 | **0.97** | **0.97** | **0.97** | **0.97** |
| MEGA | 11,845 | **0.97** | **0.97** | **0.97** | **0.97** |

Table 3.4: Imputation accuracy of minimac3 and other imputation tools. The table summarizes the accuracy of minimac3, minimac2, Impute2, and Beagle4.1 for different reference panels to impute chromosome 20. All four tools were run on 5 Mb chunks with 1 Mb overlap (13 chunks in serial or chromosome 20 yielding a total of 227,925 variants). The number of variants in the 3 MAF bins are 32,945 [$10^{-4}$-0.5%], 70,016 [0.5-5%], and 104,751 [5-50%]. The best results are highlighted in bold face.

(Table 3.2). If desired, genotypes of imputation-identified carriers can be validated through Sanger sequencing or other targeted assays prior to callback phenotyping or other follow-up analyses.

The complexity of our algorithm depends on the number of unique haplotypes in each genomic segment and the total number of such segments in the reference panel (Figure 3.1). Due to linkage between variants, the number of unique haplotypes in a genomic segment increases sublinearly with sample size. For example, in a 10kb window of simulated data, the number of unique haplotypes increased from 40 to 120 as the number of individuals increased from 200 to 1,500. Thus, minimac3 scales better than linearly; increasing the simulated reference panel from 1,000 to 20,000 individuals (20-fold) increased memory and computing time 7-fold. In real data, increasing the panel size from 1,092 to 11,455 (10-fold) increased computing time 5-fold.

### 3.3.3 Compressed File Structure

Our state-space reduction also provides an efficient way to represent haplotype data, substantially reducing file size relative to the now standard Variant Call Format (VCF). The relative efficiency of the representation reflects population genetics: for example, data from European population can be compressed more than African ancestry samples (Table D.1, p.87 in Appendix D). We adapted the VCF format to allow for these efficiencies, resulting in the **m3vcf** (minimac3 VCF) format that stores only one copy of each unique haplotype in each genomic segment. We calculated the order of disk space saved using m3vcf files compared to usual VCF files (in both unzipped and zipped formats). We found that for 1000G Phase 1 with ~1K reference samples we save ~60% of disk space when using zipped m3vcf files compared to zipped VCF files, and ~93% when compared across unzipped formats (Table D.2, p.88 in Appendix D). The saving is even greater for larger panels. For example, for the HRC reference panel with ~33K samples, we save ~84% and ~98% of disk space using zipped and unzipped m3vcf files. Our minimac3 distribution includes simple utilities to manipulate m3vcf files.

### 3.3.4 Imputation as a Service

Continued computational improvements in imputation tools are necessary for imputation to remain broadly accessible as reference panels scale to $10,000$s of sequenced genomes and $1,000,000$s of genotyped samples become available. However, a large burden of imputation and other genomics tools, is the requirement for users to master and manage large high-performance cluster jobs from the command line. To address this major need, we constructed a cloud-based imputation server. The server currently uses the minimac3 engine for rapid computation but pairs it with an accessible user interface. In the past 6 months, genomes for >1.5 million individuals have been imputed. Behind the scenes, the server divides large datasets into small chromosome segments that are processed in a parallel. Results are collected and pasted together so the process is seamless to end users.

Our imputation server also performs quality control (QC), verifying strand orientation, allele labeling, file integrity, minor allele frequency distribution, and per sample and per variant missingness. If no major problems are encountered, samples are then imputed using one of the currently available reference panels: HRC, 1000G Phase 1 and 3, HapMap Phase 2, and CAAPA. Data can be uploaded directly or by specifying a remote secure file transfer protocol (sftp) location. Data transfers are encrypted and input data are deleted after processing. As imputation proceeds, the users are provided feedback on progress, QC summary reports, email notification, and a download link for the imputed data. Our current service (`https://imputationserver.sph.umich.edu`), consisting of 12 multi-processor computers (465 cores in total) and can impute >9 million genomes using the HapMap 2 panel and >175K genomes using the HRC panel per month. Since announcing the service in October 2014, $> 7,800,000$ genomes from ~$22,000$ users have been processed.

## 3.4 Discussion

Here, we have described improved computational methods and interfaces to ensure that im-

putation remains broadly accessible - enabling researchers to rapidly process millions of samples, without first becoming experts in the intricacies of imputation software and cluster job management, and to conveniently access large reference panels of sequenced individuals, such as those from the HRC. To make this service highly scalable, we have re-engineered the core algorithms in our imputation engine. Our improved imputation algorithm based on “state-space reduction" provides a numerically faster solution for genotype imputation with no loss in imputation accuracy and enables the use of large reference panels. Our cloud-based interface simplifies analysis steps for users and can be adapted to other analysis needs, such as linkage-disequilibrium aware genotype calling from low-coverage sequence data. This trend, where cutting edge software, large data, computational power, and a friendly interface are packaged together, will become increasingly important as genomic datasets increase in size and complexity.

# CHAPTER 4

## Minimac4 - Faster Imputation through Aggressive State Space Reduction of Hidden Markov Models

### 4.1 Introduction

Despite substantially boosting the quality of imputation, modern reference panels like the HRC demand expensive computational resources. In Chapter 3, we developed minimac3 along with a web-imputation server that significantly reduced computational burden while keeping the same imputation accuracy. Despite that, the Michigan imputation server, which has over 400 cores of CPU, has a monthly throughput of ~175,000 genomes using the HRC panel. At this rate, future GWA studies which aim to have millions of samples, would need over 5 months to analyze their samples. Furthermore, the Sanger imputation server, which uses PBWT (Durbin 2014) as their imputation tool, has a higher throughput than our server, albeit at a slightly lower imputation accuracy (see Figure 2.2, p.21). Similarly, algorithmic improvements made to Beagle4.1 since its first release has made it a competing tool in terms of imputation time. In this chapter, we aim to modify our imputation algorithm to further decrease computation complexity. The boost in imputation speed would substantially increase the throughput of our imputation server as well as enable users to efficiently impute against future panels with larger number of reference samples (for example, the second release of TOPMed is predicted to have over ~80,000 samples).

In minimac3, we introduced the idea of state space reduction, which only analyzes the unique haplotypes in small genomic regions, thereby reducing the size of the Markov model state space (Das et al. 2016). To illustrate this with an example, the HRC panel has around $65,000$ haplotypes, but in small genomic segments we analyzed only ~$500$ unique templates on average. In

this chapter we implement a more aggressive form of state space reduction that merges haplotypes based on their alleles at the genotyped markers only. This further round of collapsing, along with other software engineering techniques and numerical approximations, brings additional computational savings. These new methods have been implemented in our open source `c++` package called **minimac4**.

## 4.2 Methods

Minimac4 employs a parsimonious model for the basic imputation framework based on the Li and Stephens model (described in Chapter 1, Section 1.1.1, p.6). The methods described in this chapter have three basic features. (1) We restrict the HMM model to markers that were genotyped in the study samples, which aids in a more aggressive reduction of the HMM state space and hence reduces the computational complexity. (2) We impute markers that were not genotyped using a linear approximation. These approximations negligibly reduce accuracy but significantly improve imputation speed. (3) We optimize data representation methods and improve software engineering techniques to further increase computational efficiency.

### 4.2.1 Aggressive State Space Reduction

The intuition behind imputation is to use the alleles at the genotyped markers of the study sample to find matching or similar haplotype segments from the reference panel (see Section 1.1.1, p.6). Since all the genetic information of the study sample is contained within the genotyped markers, we extended the minimac3 algorithm by collapsing haplotypes that have identical alleles only at positions that were genotyped in the sample, instead of using all the markers in a genomic segment (as in minimac3). Under the new model, reference haplotypes which are collapsed together must have the same alleles at the genotyped markers but might not agree at the intervening ungenotyped markers. We illustrate this with example in Figure 4.1, p.41. The top panel is the original state space with 9 reference haplotypes at 27 markers (as used in minimac), the middle panel is the reduced state space (minimac3), and the bottom panel is the aggressively reduced state space

(minimac4). Under the usual state space reduction, the first genomic segment yields 3 unique haplotypes. Under the above extension, the first two unique haplotypes are further collapsed as they agree at the genotyped markers (see Figure 4.1). We call this extended version "aggressive state space reduction" (ASR). Reducing the state space in this way is expected to bring a significant fold decrease in computational complexity, especially since the number of genotyped markers is only ~1% of the total number of markers. We describe some of the formulations in this chapter while rigorous details of proof are given in Appendix E. For a brief review of notations and their meaning, please refer to Chapter 3.

**Notations.** Let us label the original state space as $\boldsymbol{\mathcal{S}}_1 = \{X_1, X_2, \ldots, X_H\}$, the reduced state space as $\boldsymbol{\mathcal{S}}_2 = \{Y_1, Y_2, \ldots, Y_U\}$, and the aggressively reduced state space as $\boldsymbol{\mathcal{S}}_3 = \{Z_1, Z_2, \ldots, Z_V\}$ where $V \leq U \leq H$ are the number of templates in each of the spaces. For example in Figure 4.1, p.41, for the left most genomic block we have $H = 9$, $U = 3$ and $V = 2$. Next let us denote, for each $Z_i \in \boldsymbol{\mathcal{S}}_3$, the number of haplotypes from the space $\boldsymbol{\mathcal{S}}_2$ (and $\boldsymbol{\mathcal{S}}_1$) which were collapsed to form $Z_i$ as $m_i$ (and $M_i$). For example, for $Z_1$ we have $m_1 = 2$ and $M_1 = 6$ and for $Z_2$ we have $m_2 = 1$ and $M_2 = 3$. It is easy to observe that $\sum_{i=1}^{V} m_i = U$ and $\sum_{i=1}^{V} M_i = H$. Lastly, let us label the left probabilities for the space $\boldsymbol{\mathcal{S}}_1$, $\boldsymbol{\mathcal{S}}_2$, and $\boldsymbol{\mathcal{S}}_3$ as $L_k(.)$, $\boldsymbol{\mathcal{L}}_k(.)$ and $\mathfrak{L}_k(.)$ respectively and the right probabilities as $R_k(.)$, $\boldsymbol{\mathcal{R}}_k(.)$ and $\mathfrak{R}_k(.)$ respectively.

The first step in minimac4 is to fit the HMM on space $\boldsymbol{\mathcal{S}}_3$ to estimate posterior genotype probabilities of $Z_1, \ldots, Z_V$ only at the genotyped markers. This is done by implementing the forward equations on space $\boldsymbol{\mathcal{S}}_3$, followed by the backward equations on the same space, and then estimating the posterior probabilities. The formulas are similar to those in minimac3 with one major exception. In the aggressive version, we run the forward and backward recursions only on the genotyped markers rather than all of the markers along the chromosome. Thus, we need to re-define the transition function which should now measure the probability of a template switch across consecutive genotyped markers. A more general version of the recombination rate ($\lambda_{A:B}$) which can measure the probability of a template switch between any pair of markers $A$ and $B$ is given below (where

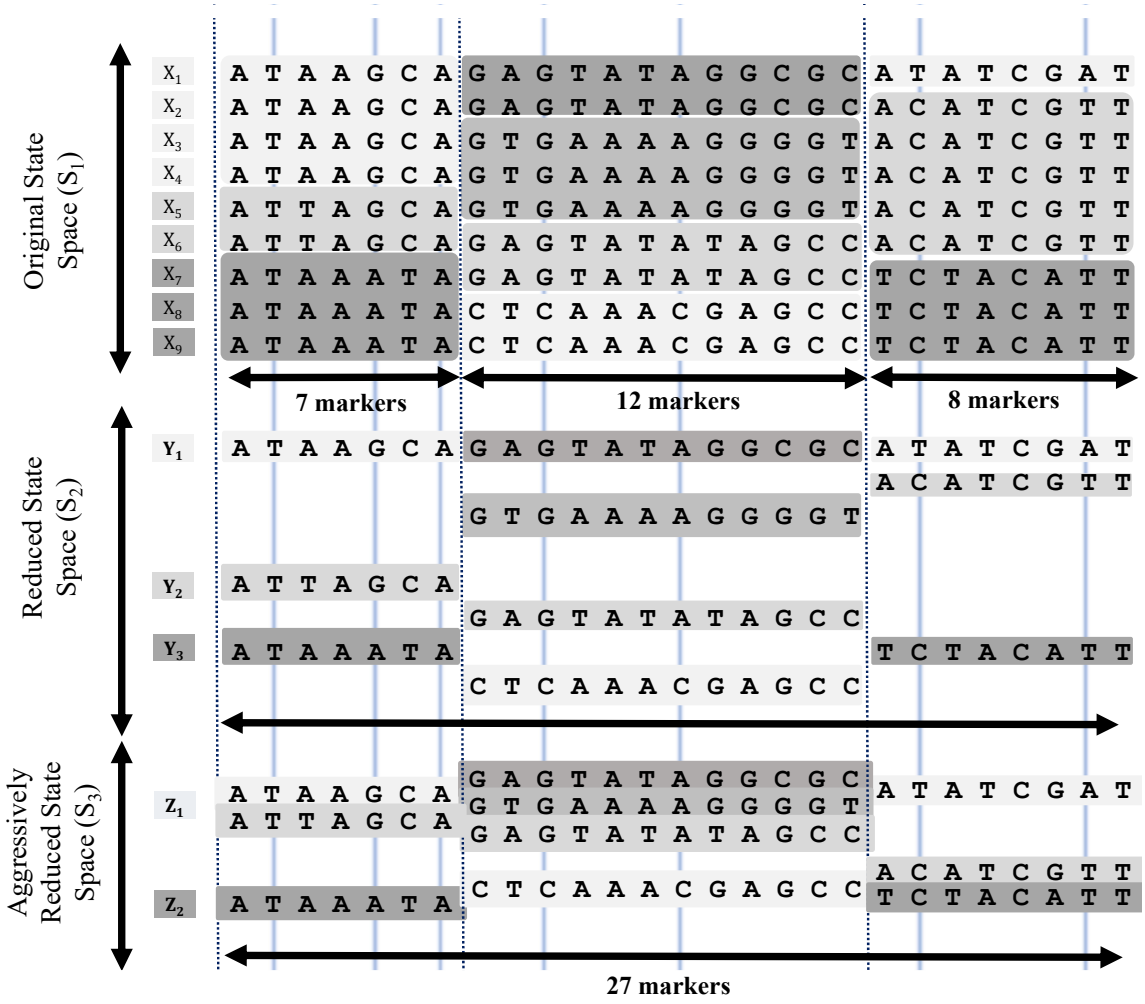


Figure 4.1: Overview of aggressive state space reduction in minimac4. We consider the first block consisting of the first 7 markers. The top panel shows the original space ($\boldsymbol{\mathcal{S}}_1$) with $H = 9$ haplotypes ($X_1, \ldots, X_9$). The middle panel shows the usual state space reduction ($\boldsymbol{\mathcal{S}}_2$) by collapsing haplotypes that are identical to each other within the first block. For example, $X_1$ to $X_4$ were collapsed to form $Y_1$. This gives 3 unique states ($Y_1, Y_2, Y_3$) for space $\boldsymbol{\mathcal{S}}_2$ ($U = 3$). The bottom panel shows the aggressive version ($\boldsymbol{\mathcal{S}}_3$) where we collapse templates from $\boldsymbol{\mathcal{S}}_2$ which match at the genotyped markers (denoted as blue vertical lines). For example, we collapse $Y_1$ and $Y_2$ to form $Z_1$ as they agree at markers 2, 5, and 7. This yields 2 states ($Z_1$ and $Z_2$) for space $\boldsymbol{\mathcal{S}}_3$ ($V = 2$). We also note that the templates collapsed together in $Z_1$ don't agree at marker 3 (ungenotyped)

$A$ and $B$ need not be consecutive and $\lambda_i$ is as defined in Section 3.2.1, p.27):

$$1 - \lambda_{A:B} = \prod_{i=A}^{B-1} (1 - \lambda_i) \tag{4.1}$$

**Forward Equations.** The first equation (4.2) transforms the left probabilities from the original space $\boldsymbol{\mathcal{S}}_1$ to the aggressively reduced space $\boldsymbol{\mathcal{S}}_3$ ($L_P \rightarrow \mathfrak{L}_P$), the second equation (4.3) gives the modified forward equation for the space $\boldsymbol{\mathcal{S}}_3$ ($\mathfrak{L}_P \rightarrow \mathfrak{L}_Q$), and the third equation (3.3) transforms the left probabilities from space $\boldsymbol{\mathcal{S}}_3$ back to the original space $\boldsymbol{\mathcal{S}}_1$ ($\mathfrak{L}_Q \rightarrow L_Q$).

$$\mathfrak{L}_P(Z_i) = \sum_{j=1}^{H} I(X_j = Z_i) \times L_P(X_j) \tag{4.2}$$

$$\mathfrak{L}_B\left(Z_i\right) = \left[\left[1 - \lambda_{A:B}\right] \mathfrak{L}_A\left(Z_i\right) + \frac{M_i \lambda_{A:B}}{H} \sum_{j=1}^{V} \mathfrak{L}_A\left(Z_j\right)\right] \times P\left(S_B | Z_i\right) \tag{4.3}$$

$$L_Q(X_j) = \mathfrak{L}_Q^R(Z_i) \times \left[\frac{1}{M_i}\right] + \mathfrak{L}_Q^{NR}(Z_i) \left[\frac{L_P(X_j)}{\mathfrak{L}_P(Z_i)}\right] \tag{4.4}$$

In the above equations:

- $A$ and $B$ are two consecutive genotyped markers.
- $\lambda_{A:B}$ denotes the template switch probability between markers $A$ and $B$ (defined earlier).
- $P\left(S_B | Z_i\right)$ denotes the genotype emission probabilities. We note that haplotypes which are collapsed together in the space $\boldsymbol{\mathcal{S}}_3$ agree at the genotyped markers and thus will have the same genotype emission probabilities. Hence, the term $P\left(S_B \,|Z_i\right)$ makes sense
- $\mathfrak{L}_Q^{NR}$ and $\mathfrak{L}_Q^R$ are the components of $\mathfrak{L}_Q$ and are defined similar to equations (3.4) and (3.5).

**Backward Equations.** Equation (4.5) transforms the right probabilities from $\boldsymbol{\mathcal{S}}_1$ to $\boldsymbol{\mathcal{S}}_3$ ($R_Q \rightarrow \mathfrak{R}_Q$), while equation (4.6) gives the modified formulation for the Baum-Welch backward equations ($\mathfrak{R}_Q \rightarrow \mathfrak{R}_P$), and the third equation (4.7) transforms back the right probabilities from $\boldsymbol{\mathcal{S}}_3$ to $\boldsymbol{\mathcal{S}}_1$

($\mathfrak{R}_P \rightarrow R_P$).

$$\mathfrak{R}_Q(Z_i) = \sum_{j=1}^{H} I(X_j = Z_i) \times R_Q(X_j) \tag{4.5}$$

$$\mathfrak{R}_A\left(Z_i\right) = \frac{M_i \lambda_{A:B}}{H}\left[\sum_{j=1}^{V} \mathfrak{R}_B\left(Z_j\right) P\left(S_B|Z_j\right)\right] + \left[1-\lambda_{A:B}\right] \mathfrak{R}_B\left(Y_i\right) P\left(S_B|Y_i\right) \tag{4.6}$$

$$R_P(X_j) = \mathfrak{R}_P^R(Z_i) \times \left[\frac{1}{M_i}\right] + \mathfrak{R}_P^{NR}(Y_i)\left[\frac{R_Q(X_j)}{\mathfrak{R}_Q(Z_i)}\right] \tag{4.7}$$

**Posterior Probabilities.** The formula below gives the posterior probability of each aggressively reduced state (4.8).

$$\begin{aligned} P\left(Z_i\,|\text{GWAS}\right) \;=\; & \left[\sum_{j=1}^{H} I(X_j = Z_i) L_P\left(X_j\right) R_Q\left(X_j\right)\right] \times \left[\frac{\mathfrak{L}_K\left(Z_i\right)}{\mathfrak{L}_P\left(Z_i\right)} \times \frac{\mathfrak{R}_K\left(Z_i\right)}{\mathfrak{R}_Q\left(Z_i\right)}\right] \\ & + \frac{1}{M_i}\left[\mathfrak{L}_K\left(Z_i\right)\mathfrak{R}_K\left(Z_i\right) - \mathfrak{L}_K^{NR}\left(Z_i\right)\mathfrak{R}_K^{NR}\left(Z_i\right)\right] \end{aligned} \tag{4.8}$$

Restricting the HMM to only genotyped markers enables us to implement the model with a lesser number of comparisons than we could in minimac3. Additionally, the HMM model fitted in minimac4 is not mathematically different from the model implemented in minimac3, in the sense that it generates the exact same posterior probabilities at the genotyped markers as they would in minimac3. Refer to the Appendix E for proof and further details.

### 4.2.2 Imputation of Ungenotyped Markers

The second step involves imputing the ungenotyped markers. In minimac4, instead of implementing the HMM across all markers along the chromosome, we decided to use constant probabilities for a small ungenotyped region. To be more specific, for each genotyped marker we define a 'flanking region' surrounding that marker. These flanking regions are non-overlapping and assembled sequentially by forming intervals using the mid-points that lie between two consecutive

genotyped markers (see Figure 4.2). For each such flanking region, we use the posterior probabilities estimated at the corresponding genotyped marker to impute the ungenotyped markers in that region.

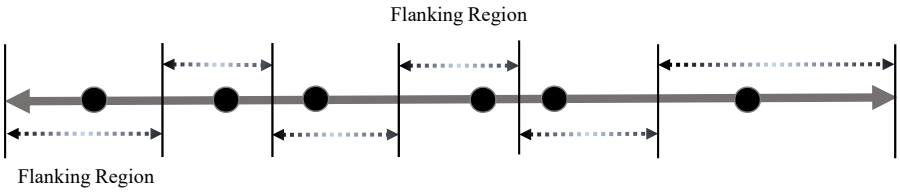


Figure 4.2: Assembly of flanking regions for imputation. The figure describes the formation of the flanking regions. The solid black circles represent the genotyped markers. The vertical black lines denote the mid-points between consecutive genotyped markers. These lines form the end-points for the flanking regions. Each genotyped marker lies inside its own flanking region, but need not be at the mid-point of that region

However, as mentioned previously, haplotypes which have been collapsed together in space $\boldsymbol{\mathcal{S}}_3$ might not agree at all the intermediate ungenotyped markers. Thus, we need to transform the probabilities from the space $\boldsymbol{\mathcal{S}}_3$ to the space $\boldsymbol{\mathcal{S}}_2$ and use the transformed posterior probabilities to impute the ungenotyped markers in the flanking region. To explain the notations, let $K$ be a genotyped site and $Y_j \in \boldsymbol{\mathcal{S}}_2$ be a template that was collapsed to form $Z_i \in \boldsymbol{\mathcal{S}}_3$. The crude way to transform probabilities from $\boldsymbol{\mathcal{S}}_3$ to $\boldsymbol{\mathcal{S}}_2$ would be to first transform from $\boldsymbol{\mathcal{S}}_3$ to $\boldsymbol{\mathcal{S}}_1$ using equation (4.4) and then from $\boldsymbol{\mathcal{S}}_1$ to $\boldsymbol{\mathcal{S}}_2$ using equation (3.1). However, the following equation directly transforms them from space $\boldsymbol{\mathcal{S}}_3$ to space $\boldsymbol{\mathcal{S}}_2$, where $N_j$ is number of templates from space $\boldsymbol{\mathcal{S}}_1$ collapsed to form $Y_j$ (defined earlier in Section 3.2.1). This equation eliminates the need of space $\boldsymbol{\mathcal{S}}_1$ as an intermediate step and this is beneficial in reducing complexity, especially since the cardinality of $\boldsymbol{\mathcal{S}}_1$ is much higher than that of $\boldsymbol{\mathcal{S}}_2$ or $\boldsymbol{\mathcal{S}}_3$. A similar equation can be derived for the right probabilities which can then be combined with the left probabilities to yield the final posterior probabilities for the templates in space $\boldsymbol{\mathcal{S}}_2$. Detailed proofs are found in Appendix E.

$$\boldsymbol{\mathcal{L}}_Q(Y_j) = \boldsymbol{\mathfrak{L}}_Q^R(Z_i) \times \left[\frac{N_j}{M_i}\right] + \boldsymbol{\mathfrak{L}}_Q^{NR}(Z_i) \left[\frac{\boldsymbol{\mathcal{L}}_P(Y_j)}{\boldsymbol{\mathfrak{L}}_P(Z_i)}\right] \tag{4.9}$$

**Most Likely Templates.** In our internal studies with large reference panels, we observed that for a specific sample in a specific genomic region, only a handful of haplotype templates have a high posterior probability (say $> 0.01$) (see Figure E.1 in Appendix E, p.93). Analyzing only the top few most probable templates, instead of all of them, can further reduce computational time. When fitting the HMM at the genotyped markers, we create a set of 'most likely templates' from space $\boldsymbol{\mathcal{S}}_3$ (states with posterior probability more than a pre-determined threshold). When we impute the ungenotyped markers, we only transform the probabilities of these most likely templates from $\boldsymbol{\mathcal{S}}_3$, instead of all the templates $Z_1, \ldots, Z_V \in \boldsymbol{\mathcal{S}}_3$. The final number of templates from space $\boldsymbol{\mathcal{S}}_2$ to be used for imputing the ungenotyped markers, is thus, much less than $U$ (the total number of templates in space $\boldsymbol{\mathcal{S}}_2$).

### 4.2.3 Software Engineering Techniques

Apart from the above algorithmic developments, we have also leveraged other software engineering techniques. Compact representation of data not only reduces memory requirements, but also boosts computation speed by enabling faster search and data access.

**Compact Reference Panel.** In minimac3, we described the **m3vcf** format for storing the reference panel, that only encoded the unique haplotype templates in small genomic regions. This format significantly reduced disk space requirements for the data as well as time to load the data in the physical memory. In minimac4, we extend these efforts, by exploiting the fact that rare variants, despite making up a significant percentage of the total number of variants, exhibit only a few copies of the minor allele in the data. Consequently, for all markers with a MAF less than a pre-determined threshold, only the indices of the haplotypes which carry the minor allele are stored. These indices are stored in a sorted data structure and can be efficiently retrieved using a binary search technique. If the index of a sample haplotype does not belong to that list, we know that it carries the reference allele. The compression strategy also boosts imputation speed since im-

puting these rare markers requires many fewer comparisons (e.g. only $3$ comparisons are needed, instead of $100$, to impute a marker with 3 minor alleles, when there are $100$ unique templates in that genomic segment).

The second strategy was to write binary files. This was essential because, as we keep increasing size of the reference panel, the time required to read data from the disk starts to dominate the time required for imputation. This might be because, reading the data from disk is always linear in the number of reference haplotypes, while the minimac3 algorithm scaled better than linear (see Table 3.3, p.33 in Chapter 3). The downside of writing binary files is that they are no longer human readable.

**Memory Efficient Data Structures.** Compact representation of the imputed allele probabilities also decreased the memory requirements. Though the HMM calculations are performed using 4-byte floating point variable to prevent rounding errors, the final imputed value is stored as a 1 byte value. To achieve this, for a fixed haplotype dosage value between 0 and 1, we find the nearest element in set $A = \left\{ \frac{2i+1}{512} : i \in 0, 1, \ldots, 255 \right\}$ and approximate that imputed value by that element. The maximum error introduced this way is $\frac{1}{512} \approx 0.00195$ (less than $0.2\%$). Since there are 256 elements in the set $A$, we can store them using a single byte value (compared to 4 bytes for floating point variable) thus giving us a 4-fold memory efficiency on some data structures.

**Multi-threading.** We greatly improved the parallelization module by using modern compiler directives and library routines from the OpenMP® programming interface. Some of the key features implemented are described here. (1) We reduced bottlenecks due to synchronization between multiple threads by reducing the number of unnecessary steps in critical segments[1]. (2) Critical segments that can run simultaneously were given different construct names to allow them to execute at the same time. For example, creating summary statistics and writing the output file are both individually critical steps but can be run in parallel since neither step hinders the other. (3) We used

[1]Segments of code that should be executed only thread at a time. For example, writing the output file is a critical step since two threads trying to write on the same file simultaneously would give erroneous results.

in-built data sharing and initializing clauses (e.g. `shared`, `threadprivate`) to improve speed of data synchronization across threads and thus reduce bottlenecks arising from repeated memory reallocations. (4) The default output format (VCF) requires the final imputed data to be transposed (since VCF format stores markers along rows and samples along columns). Additionally, multi-threading compels samples to be imputed in a random order which later needs to be sorted back; this further adds to the time spent in the critical step of file output. We implemented a system of nested parallelization that removed these bottlenecks due to file output. A pre-fixed number of consecutive samples are imputed in parallel together. The resulting imputed data from such a batch is stored in the physical memory and then written in a partial VCF file with the samples in the correct order. Once all such batches are finished the partial VCF files are merged back. This step is greatly beneficial because it allows the next batch of samples to start imputing while the previous batch is still being written on disk, thus removing the bottleneck.

## 4.3 Results

### 4.3.1 Comparison of Imputation Tools

We compared the imputation run time and physical memory required by minimac4, to minimac3, Beagle4.1 and Impute2. We imputed into 1,000 GWA samples using typical reference panels: 1000G Phase 1, 1000G Phase 3, the HRC, and the TOPMed panel. The number of non-missing genotypes provided by the panels are ~0.67 billion, ~5.2 billion, ~57.4 billion, and ~176.2 billion respectively. The results are shown in Table 4.2, p.49. Estimates of run-time for the whole genome were interpolated from analysis on chromosome 20. For minimac3 and minimac4 we used the **m3vcf** format for the reference panel, for Beagle4.1 we used the **bref** format, and for Impute2 we used the Oxford **hap/legend** format. The m3vcf and bref are compact representations of the haplotype data and thus aid in faster reading of the data. When imputing against the TOPMed panel, minimac4 is ~2.5 times faster than minimac3 and ~3 faster than Beagle4.1. However, the memory consumed by minimac4 is slightly more than Beagle4.1 (~0.74 Gb versus ~0.7 Gb). For HRC panel it is ~2.5 times faster than Beagle4.1 and the memory consumed is almost a third (~0.68

Gb versus ~1.78 Gb, respectively)

Minimac4 scales quite well with increasing panel size and marker density. For example, the number of non-missing genotypes in TOPMed is ~3 times that of HRC, but the time taken for imputation is quite similar (~290 hours versus ~276 hours). Similarly, the time taken for 1000G Phase 3 is less than twice as that of Phase 1 (~56 hours versus ~30 hours, respectively) although the Phase 3 panel provides ~7 times more genetic information. Between minimac3 and minimac4, the fold increase in speed is less in HRC compared to TOPMed and this is probably because HRC panel was created at an MAC greater than 5, while the TOPMed panel starts at MAC $\geq 2$, thus enriching the rarer spectrum.

### 4.3.2 Comparison of Imputation Accuracy

We next examined the change in imputation accuracy due to approximations implemented in minimac4 (using most likely templates and constant probabilities in ungenotyped regions). We compared the accuracy of minimac4 to impute 10 CG samples using 4 reference panels to that of minimac3 (reference panels used are 1000G Phase 1, 1000G Phase 3, HRC, and TOPMed). The results are shown in Table E.1, p.94 in Appendix E. For a comparison of accuracy between minimac3 and other methods see Table 3.4, p.34 in Chapter 3. The fall in accuracy due to approximations in minimac4 is negligible; the maximum fall in aggregate[2] $r^2$ across all AF bins is ~0.002 for the TOPMed panel, and ~0.001 for the HRC panel. The difference in accuracy also decreases as the panel size increases from 1000G Phase 1 to TOPMed (maximum difference in 1000G Phase 1 is ~0.049). The larger panels improve the chances of finding shared templates and thus mitigate the effect of numerical approximations applied to the HMM.

### 4.3.3 Comparison of Reference Panels

We estimated the average run time for different reference panels using minimac4. We imputed into 1,000 GWA samples on chromosome 20 against 1000G Phase 1, 1000G Phase 3, HRC, and

[2]See footnote on page 20 for definition of aggregate $r^2$

| Reference Panel | # Genotypes | # Samples | # Variants | minimac4 | minimac3 | Impute2 | Beagle 4.1 |
|---|---|---|---|---|---|---|---|
| Time (in CPU-hours) | | | | | | | |
| 1000G Phase 1 | ~0.67B | 1,092 | ~617K | **29** | 49 | 340 | 50 |
| 1000G Phase 3 | ~5.2B | 2,504 | ~1.04M | **56** | 112 | 980 | 120 |
| HRC | ~57.4B | 32,470 | ~884K | **276** | 414 | 1,351 | 728 |
| TOPMed | 176.2B | 19,815 | ~4.44M | **290** | 783 | 60,500 | 900 |
| Memory (in CPU-Gigabytes) | | | | | | | |
| 1000G Phase 1 | ~0.67B | 1,092 | ~617K | **0.07** | 0.09 | 0.91 | 0.51 |
| 1000G Phase 3 | ~5.2B | 2,504 | ~1.04M | **0.09** | 0.13 | 1.88 | 0.56 |
| HRC | ~57.4B | 32,470 | ~884K | **0.68** | 0.95 | 22.08 | 1.78 |
| TOPMed | 176.2B | 19,815 | ~4.44M | 0.74 | 1.24 | 107.7 | **0.70** |

Table 4.2: Computational requirements of minimac4 and other imputation tools. This table summarizes the running time and memory requirements of minimac4, minimac3, Impute2, and Beagle4.1 for different reference panels to impute 1,000 whole genomes (running time interpolated from analysis on chromosome 20). All four tools were run on 5 Mb chunks with 1 Mb overlap (13 chunks in serial or chromosome 20 yielding a total of 227,925 variants). Minimac4, minimac3, and Impute2 were run with pre-calculated estimates. Minimac4, minimac3, and Beagle4.1 were run with their respective input file formats (m3vcf and bref respectively) while Impute2 was run on Oxford format files. The best results are highlighted in bold face.

TOPMed and the results are shown in Table 4.1, p.50. The main segments of the minimac4 algorithm can be categorized as follows: (a) reading the reference data from disk, (b) re-compressing the reference data (from space $\boldsymbol{\mathcal{S}}_2$ to $\boldsymbol{\mathcal{S}}_3$), (c) imputing the study samples, and (d) writing the imputed data to the disk. The time required in segments (a) and (b) are independent of the number of GWA samples and contribute to an overhead cost. Segment (c) is the only section that benefits from multi-threading. For example, multi-threading reduces the imputation time (not total run-time) by 5-fold and 10-fold when using 5 and 10 CPUs respectively, when compared to a single-thread process (see Table 4.1). Another interesting feature in Table 4.1: when using 10 CPUs to impute against the TOPMed panel, the time required to write the imputed data to disk (which should be in serial) is almost equal to the time taken to impute the samples (which has been parallelized).

| **Time (mins)** | **1000G Phase 1** | **1000G Phase 3** | **HRC** | **TOPMed** | **TOPMed (5 cpus)** | **TOPMed (10 cpus)** |
|---|---|---|---|---|---|---|
| File Read | 0.5 | 0.8 | 2.2 | 4.0 | 4.1 | 4.1 |
| File Write | 10.6 | 17.5 | 11.9 | 38.0 | 39.0 | 38.8 |
| Re-Compression | 0.1 | 0.3 | 5.0 | 3.4 | 3.8 | 3.9 |
| Imputation | 24.5 | 51.8 | 312.0 | 302.8 | 62.5 | 31.5 |
| Total Run Time | 35.7 | 70.4 | 331.1 | 348.2 | 109.4 | 78.3 |
| Speed Up (compared to minimac3) | 1.7x | 2.0x | 1.5x | 2.8x | 3x | 3.7x |

Table 4.1: Comparison of imputation time for different reference panels. This table summarizes the amount of time consumed in different steps during imputation against some typical reference panels. In this table we describe the time profiles for the 1000G Phase 1, 1000G Phase 3, HRC, and TOPMed to impute 1,000 samples on chromosome 20. For the TOPMed panel we also compare the time profile required by a single thread process to multi-thread processes using 5 and 10 CPUs.

The results signify that using more than 10 CPUs would not benefit users beyond a limit, since the bottleneck of the process would then become the file I/O and that cannot be parallelized. This feature is a direct application of Amdahl's law (Amdahl 1967) which states that theoretical speedup is always limited by the part of the task that cannot benefit from the multi-threading improvement. For example, using 10 CPUs on the TOPMed panel decreased the total run time by only ~4.4 fold

(~78 minutes versus ~348 minutes). This is because it takes ~40 minutes to write the imputed data to disk, which stays constant irrespective of the number of CPUs used (see Table 4.1)

The results from Table 4.1 have been summarized to predict the total run-times for different panels using minimac4 and have been documented in Table 4.3, p.51. It is estimated to take ~6 hours to impute 1,000 samples on chromosome 20 using the HRC panel (~7.2 minutes of overhead and $1,000 \times 21$ seconds to impute $1,000$ samples). We note that the 'time per sample' also includes the time taken to write the imputed data to disk and thus these numbers from Table 4.3 cannot be transformed to obtain multi-thread estimates.

| **Time for Chr 20** | **1000G Phase 1** | **1000G Phase 3** | **HRC** | **TOPMed** |
|---|---|---|---|---|
| Overhead | 0.6 min | 1.1 min | 7.2 min | 7.5 min |
| Time per Sample | ~2 secs | ~4 secs | ~21 secs | ~20 secs |

Table 4.3: Predicted run-time for different reference panels. This table provides an estimate of total run-time on a single CPU to impute against the major reference panels, namely, the 1000G Phase 1, 1000G Phase 3, HRC, and TOPMed. Note that the time per sample includes the time to write the imputed data to disk and these numbers cannot be transformed to obtain multi-thread estimates. See Table 4.1 for time taken in a multi-thread process.

We also compared the accuracy in European samples when imputed against the above 4 reference panels. We measured aggregate $r^2$ after imputing into 10 CG samples and the results are shown in Figure 4.3, p.52. TOPMed provides the highest accuracy for studies of European ancestry, increasing the $r^2$ to ~$64\%$ (the $r^2$ for HRC is ~$59\%$ and for 1000G Phase 3 is ~$30\%$).

### 4.3.4 Sensitivity to Parameter Values

Minimac4 has been enabled with a default system of automated chunking where it analyzes the data sequentially through non-overlapping chunks. The main benefit of imputing chunks as opposed to imputing whole chromosome is lower memory consumption; this in turn enables more samples to be imputed in parallel. The main drawback is that chunking breaks down long haplotype

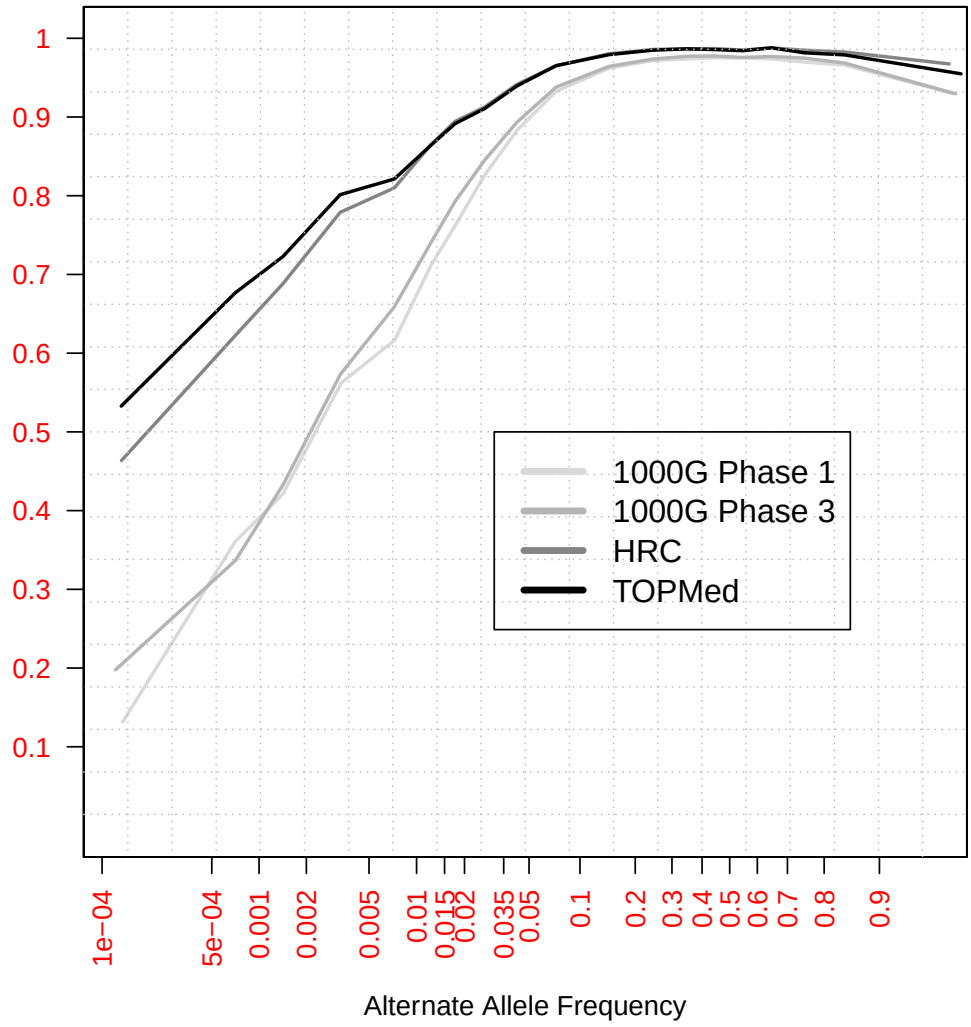


Figure 4.3: Comparison of 1000G Phase 1, Phase 3, HRC, and TOPMed. The plot shows the imputation accuracy (aggregate $r^2$) for imputing into 10 CG samples using 1000G Phase 1, 1000G Phase 3, HRC, and TOPMed reference panels.

segments, which might interfere with the haplotype matching process and consequently reduce imputation accuracy, especially for variants along the edges of the chunk. The decrease should would also be significant for rarer variants since longer segments of haplotypes might be needed to accurately impute them. Additionally, overlapping regions between chunks ends up being imputed twice and adds to the total run-time. In Figure 4.4, p.53, we plot the imputation accuracy for imputing into 100 GWA samples using the HRC panel on chromosome 20. We evaluated the effect of different chunk lengths ranging from 0.5Mb to 30Mb. While smaller chunk lengths (less

than 5Mb) decrease the accuracy substantially, large chunk lengths (greater than 15Mb) yield the same accuracy as chromosome wide imputation. The default value in minimac4 is to analyze chunks of length 20Mb with a buffer region of 3Mb on either side of the chunk; users have the option to change these values. In our internal studies, we have found that these default values to be sufficiently large enough to provide maximum accuracy across all reference panels and still be beneficial in reducing memory consumption. For example, the default values reduce memory consumption ~12 fold and increase total run time by ~20%, when imputing chromosome 2 (with no fall in imputation accuracy).

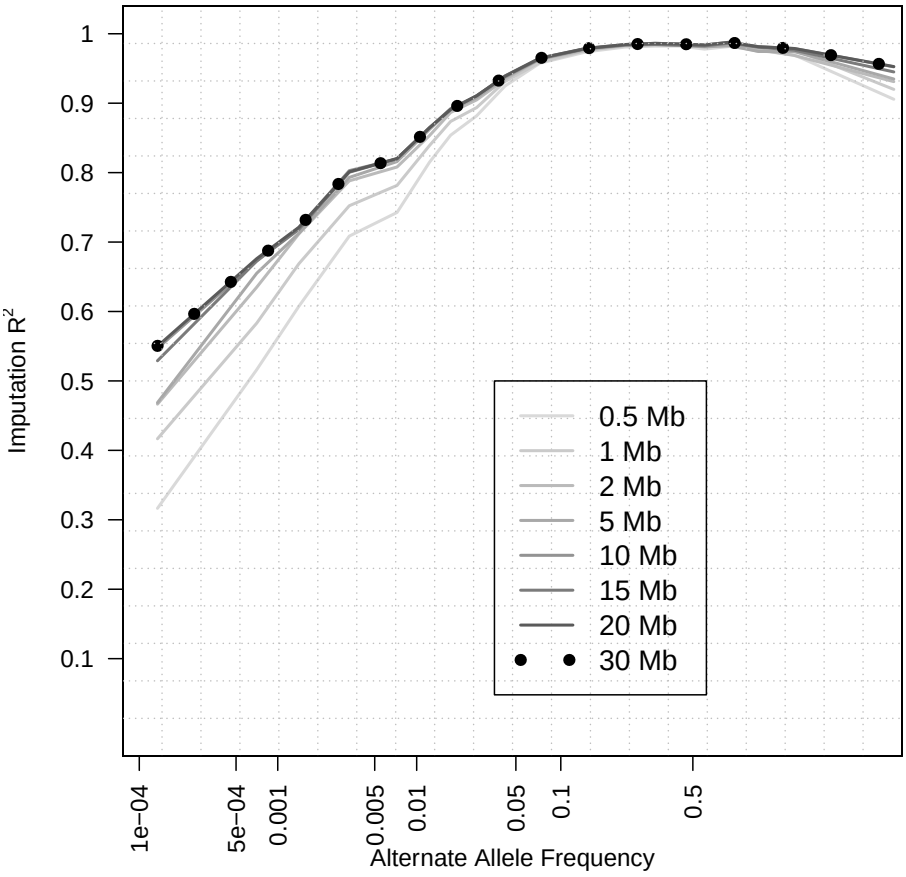


Figure 4.4: Effect of chunking on imputation accuracy. The figure describes the effect of different chunk sizes on the imputation accuracy. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$. The gray lines represent accuracy profiles for different values of chunk length used in minimac4. Chunk lengths smaller than 5Mb decrease the accuracy substantially while those larger than 15Mb yield the same accuracy as chromosome wide imputation. The default value in minimac4 is to analyze chunks of length 20Mb with a buffer region of 3Mb on either side of the chunk

### 4.3.5 Imputation in Non-European Samples

We also evaluated the accuracy administered by different reference panels for samples of African, admixed American, East Asian and South Asian ancestry. We imputed into 10 GWA samples of each of the above ancestry. The study samples were extracted from the 1000G Phase 3 reference panel, and all markers except those on the Illumina® Duo 1M Chip were masked to mimic a typical GWA study. We imputed into these samples using 1000G Phase 3, HRC, and TOPMed reference panels and then calculated aggregate imputation $r^2$. For African and admixed American ancestry we found that the TOPMed reference panel provides significantly higher accuracy than 1000G or HRC panels (see Figure H.1, p.101 and Figure H.2, p.102 in Appendix H). For example, for African samples the TOPMed panel provides ~63% aggregate $r^2$ while the HRC panel provides ~54% aggregate $r^2$ at variants with MAF ~0.1%. However, samples of East Asian and South Asian benefit the most from 1000G Phase 3 panel (see Figure H.3, p.103 and Figure H.4, p.104 in Appendix H). 1000G Phase 3 yields aggregate $r^2$ of ~55% compared to ~33% accuracy provided by TOPMed for South Asian samples at variants with MAF between 0.5% and 1%. The results demonstrate that the TOPMed panel is consequently the choice reference panel for samples of European, African and Admixed American ancestry while 1000G Phase 3 should be for East Asian and South Asian samples.

## 4.4 Discussion

In this chapter we presented some methods on improving the speed of imputation algorithms while still maintaining similar accuracy. The most effective way to gain imputation accuracy is to increase the size of the reference panel. Consequently, to keep imputation with larger panels a computationally feasible practice, imputation algorithms need to constantly improve themselves. Although, due to unavailability, we have not been able to test minimac4 on reference panels with millions of samples, we expect the algorithmic developments to scale quite well. The main reason behind this expectation being: the number of markers on a genotyping chip is far less than the

total number of markers and does not change with increasing number of samples or markers in reference panel; hence we expect the aggressively reduced space to stay much smaller than the original reduced space. For instance, the HRC panel has ~$66,000$ haplotypes but on an average we analyzed ~$500$ unique templates in minimac3 and ~$80$ unique templates in minimac4.

The results from this chapter highlight certain salient features in next generation imputation studies. First, the study on non-European samples don't agree with the idea that using a larger panel always increases the imputation accuracy. The HRC panel properly contains all samples from 1000G Phase 3 but study samples of Asian ancestry benefit more from the 1000G panel. This might be because the HRC panel is predominantly European (the only non-European samples in HRC are contributed by the 1000G) and phasing the samples of HRC jointly depletes the contribution of haplotypes unique to Asian ancestry. The haplotypes of the Asian samples in HRC show more resemblance to European haplotypes, thus reducing chances of finding good matches with Asian study samples. This could also be the reason why TOPMed panel provides better accuracy for African and Admixed American samples but the least accuracy for South and East Asian samples, as the TOPMed panel is highly enriched with African samples but has poor representation from Asian samples. Such artifacts are also seen in the field machine learning, especially in algorithms that rely on a fundamental assumption that the training and test data are drawn from identical distributions. A future direction of imputation algorithms could be to develop methods of transfer learning that can control for such 'negative transfer' of information.

Second, imputation accuracy is not affected by running analysis on chunks instead of the whole chromosome, provided the chunks are large enough. In all our experiments, chunks of 15 Mb length with around 3 Mb buffer on either side have been sufficiently large. Third, future imputation tools should emphasize improving the efficiency of file I/O as these steps will soon become main bottlenecks of imputation analysis. For large GWA studies with millions of samples, we might suggest saving only the most-likely genotypes (instead of a continuous imputed dosage). This way users can download m3vcf format for their imputed data which would ease the process of file transfer/sharing. Moving to hard call genotypes from continuous dosages might slightly reduce

the power of association studies, but the ease with which hard genotypes can be compressed would mitigate the fall in power, especially when millions of samples are being tested together. We are currently developing toolboxes that can perform association tests directly on the m3vcf format. These tools would additionally increase the speed of association studies.

As variant density in commercial chips increases, the fold increase in speed of minimac4 is expected to decrease compared to minimac3. While this is a genuine concern, based on current trends we expect future reference panels to have a more detailed catalog of structural variants which would also increase the number of variants in the reference panels. As long as the proportion of genotyped variants is small, we expect minimac4 to produce computational savings compared to minimac3. However, the main caveat of minimac4 lies in the fact that it has an overhead cost and hence it may not be beneficial for GWA studies with small number of samples. As an example, it takes 20 seconds per sample to impute against the TOPMed panel on chromosome ~20, while the overhead cost is ~8 minutes. Thus, it might not make sense to impute a study of 10 samples as the overhead would be much larger compared to actual imputation time (~4 minutes). For such instances of studies with small number of GWA samples (less than 100), we recommend using minimac3.

# CHAPTER 5

# MetaMinimac - A Simple and Flexible Method to Combine Imputed Data from Multiple Reference Panels

## 5.1 Introduction

Genotype imputation has become an integral tool in human gene association studies. Imputation is an in-silico method in which a reference panel is used to infer genetic variants that were not genotyped in a study sample. In Chapter 1, Section 1.1.1, we outlined how algorithms for genotype imputation require a reference panel of densely sequenced genotypes to be used as templates for copying genetic information. In the last decade, apart from public projects like the HapMap Project or 1000G Project, other large-scale sequencing studies have designed datasets to be used as reference panels in their respective imputation studies. Examples include a study on an isolated island in Italy, Sardinia (Francalacci et al. 2013), a study on indigenous Dutch populations (GoNL 2014), and a study on Icelandic samples (Gudbjartsson et al. 2015). To exploit this ever-increasing abundance of information, it is desirable to combine the available reference sets, as this is likely to increase the chance of imputing lower frequency and rare variants more accurately. However, how to best construct a combined panel from the rapidly accumulating reference datasets from around the world, is still a challenging question.

### 5.1.1 Motivation

The first hurdle in merging different reference panels is to determine the final set of variants, since typically different studies will have different sets of variants genotyped. One simple solution would be to choose the subset of variants that are present in all the datasets. However, this reduces

the total number of variants in the merged data and might make it difficult to identify shared haplotype segments. An ideal solution would be to consider the union set of variants, and jointly call the variants in all the samples from their respective sequence alignment files (to be referred to as the “joint method"). The HRC reference panel was created using this approach. Still, variant calling is a highly compute intensive job and isn't always a feasible solution for merging reference panels.

An easy substitute for variant calling would be to combine the reference panels by first treating them as reference panels for each other and then “cross imputing” the missing variants into each panel. The cross-imputed panels would have the same set of variants and can thus can be easily merged. However, cross-imputation followed by imputation might decrease statistical certainty in downstream association studies. This approach has been used in Impute2 and was found to be neither helpful nor harmful when compared to a large, population-specific reference panel (Huang et al. 2015). Furthermore, certain reference panels like the HRC and TOPMed are not available publicly for download and hence cannot be merged with in-house reference panels. Merging panels also requires researchers to identify and remove overlapping samples across studies in order to prevent bias in the imputation estimates. Lastly, by adding more haplotypes and variants to a reference set, a considerable higher computational burden is imposed on imputation.

The accuracy of genotype imputation depends on the following key factors: density of genotyping chip, haplotypic diversity of the study population, the number of samples in the reference panel, and genetic similarity between the reference panel and study samples (Huang et al. 2009). To demonstrate the effect of some of these factors, we imputed into 20 GWA study samples from a Norwegian population (Krokstad et al. 2012, HUNT study) using the following reference panels: 1000G Phase 1 with ~1,000 individuals, 1000G Phase 3 with ~2,000 individuals and an internal HUNT WGS data with ~1,000 individuals. While increasing the number of reference samples from 1000G Phase 1 to 1000G Phase 3 yields a clear increase in accuracy, using a more genetically similar reference panel like HUNT gives a much steeper rise at the same number of reference samples (Figure 5.1, p.59). Previous studies have also reported that a modestly sized “internal” reference

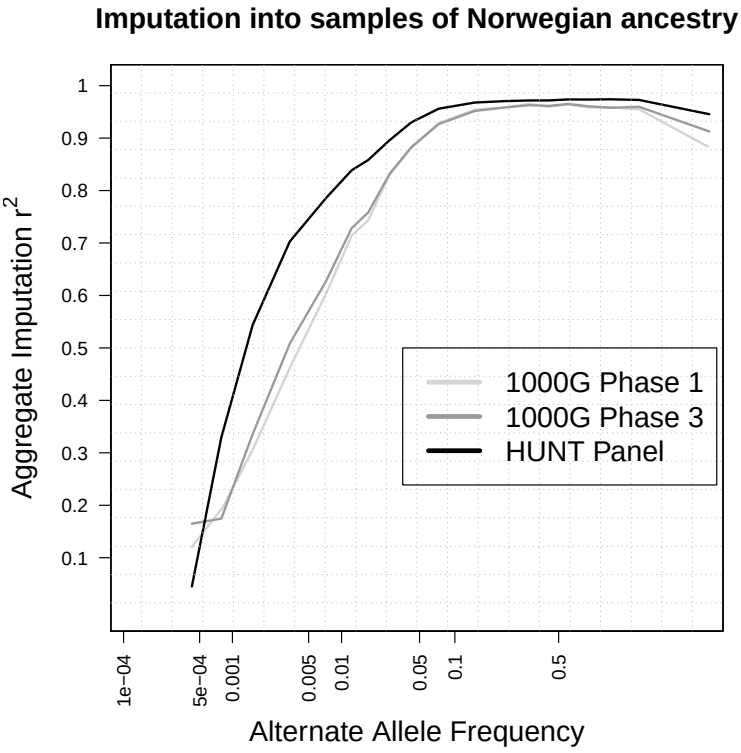


Figure 5.1: Imputation into 20 Norwegian samples. The results show minor improvement with 1000G Phase 3 but are much better with HUNT WGS panel (compared to 1000G Phase 1)

panel of genetically more similar samples yields greater imputation accuracy than a larger "external" panel from a different population, even if the divergence time between the two populations is small (Deelen et al. 2014; Jewett et al. 2012; Pistis et al. 2014).

In this chapter, we introduce the idea of meta-imputation, which is a simple and flexible approach that integrates imputed data from multiple reference panels without interfering with the original imputation algorithm or requiring access to the reference data. The term meta-imputation has been used analogous to the term 'meta-analyses' for variant association. In meta-analysis, we combine summary statistics obtained from different GWA studies that are imputed against the same reference panel. In meta-imputation, we combine imputed data for the same GWA study imputed against different reference panels. Meta-analysis is beneficial for improving power for association test while meta-imputation is beneficial for increasing accuracy of rare variant imputation. Additionally, meta-imputation is a solution to the challenges faced in merging reference panels.

## 5.2 Methods

Meta-imputation works by separately imputing the genotyped sample against different references, and then combining the inferred genotype likelihoods at the overlapping markers, using imputation-dependent scores in a weighted average function. We have implemented our method in a publicly available `c++` package called **metaMinimac**.

### 5.2.1 Model Description

Let us assume the most general set up where there are $K$ reference panels denoted as $\boldsymbol{P}1, \boldsymbol{P}2$, ...,$\boldsymbol{P}K$. We can assume that all the panels have the same set of markers (denote $M$), since markers which are specific to only one reference panel, if any, don't gain any genetic information from the other panel and thus don't need to be meta-imputed. Let us denote the subset of variants which were genotyped as $\boldsymbol{G} \subset \{1, \ldots, M\}$. We describe methods for meta-imputing a single sample haplotype. Let us denote for a study haplotype, the estimated haplotype dosage from the $k^{th}$ reference panel as $X_{k,1}, \ldots, X_{k,M}$ (for each $k = 1, \ldots, K$). We denote the meta-imputed haplotype dosage at each marker as $Z_1, \ldots, Z_M$ which is represented as a weighted function of the individual dosages (Eq. 5.1):

$$Z_i = \sum_{k=1}^{K} w_k X_{k,i} \tag{5.1}$$

where the weights $w_k$ satisfy $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$ and would be determined by an ad-hoc method that utilizes imputation dependent accuracy scores. The weights indicate which reference panel to give preference to while determining the final meta-imputed dosage at each marker. A panel that provides better matches of shared haplotypes with the GWA haplotype would have a higher weight in the above function. The weights are allowed to vary based on the marker position since different panels might be better depending on the region of the genome.

### 5.2.2 Leave-One-Out Imputation

In a hypothetical situation, if we knew the genotypes of the GWA sample at ungenotyped

markers, we could compare these genotypes to the imputed data from each reference panel and consequently get estimates of the weights. In real scenarios, we mimic this approach by comparing the genotypes at the genotyped markers to the imputed dosages from each panel obtained by hiding data for each genotyped marker in turn. While estimating dosages at genotyped markers, in addition to reporting the usual imputed dosage, **minimac3/4** also reports a leave-one-out (loo) dosage which is estimated by masking the genotyped alleles of the study sample at that marker and then imputing it back.

**Example.** In Figure 5.2, p.62 we give a simple example of leave-one-imputation using a single panel of 6 reference haplotypes ($X_1, \ldots, X_6$) and $M = 10$ markers, where only 3 markers were genotyped ($G = 1, 5, 9$). The top left-panel describes the regular imputation, where 2 haplotypes ($X_3$, $X_5$; shaded in bold) match the GWA haplotype ($S_G$) based on the three genotyped positions. The remaining three panels explain the leave-one-out imputation for each genotyped marker in $G = \{1, 5, 9\}$. In the top-right panel, when we hide the observed allele for $S_G$ at marker 5, we note that 3 haplotypes ($X_2$, $X_3$, $X_5$) match the study haplotype (based on marker 1 and 9). Consequently, the loo dosage at marker 5 would be $(1 + 0 + 0)/3 = 0.33$ (while the usual impute dosage was $0.0$). In the bottom-left panel, hiding the observed allele at marker 1 yields the same haplotypes matching the study sample ($X_3$, $X_5$) (based on marker 1 and 9). Thus, the loo dosage at marker 1 stays same as the usual dosage ($= 1.0$). Similarly, in the bottom-right panel, when we hide the allele at marker 9, we have 5 reference haplotypes matching the study sample ($X_1$, $X_3$, $X_4$, $X_5$, $X_6$; based on marker 1 and 5) yielding a loo dosage of $0.4$ (usual imputed dosage was $1.0$).

At each genotyped marker ($i \in \boldsymbol{G}$), let us denote the loo dosage from the $k^{th}$ panel at that marker as $\tilde{X}_{k,i}$ (the respective usual dosages are denoted as $X_{k,i}$). Thus, for the sample shown in Figure 5.2 and the given reference panel, we have $(X_1, X_5, X_9) = (1, 0, 1)$ and $(\tilde{X}_1, \tilde{X}_5, \tilde{X}_9) = (1.00,\ 0.33,\ 0.40)$. Since the loo dosage has been estimated by hiding observed genotype information from a marker, we can compare them to the observed genotypes to get an idea of the

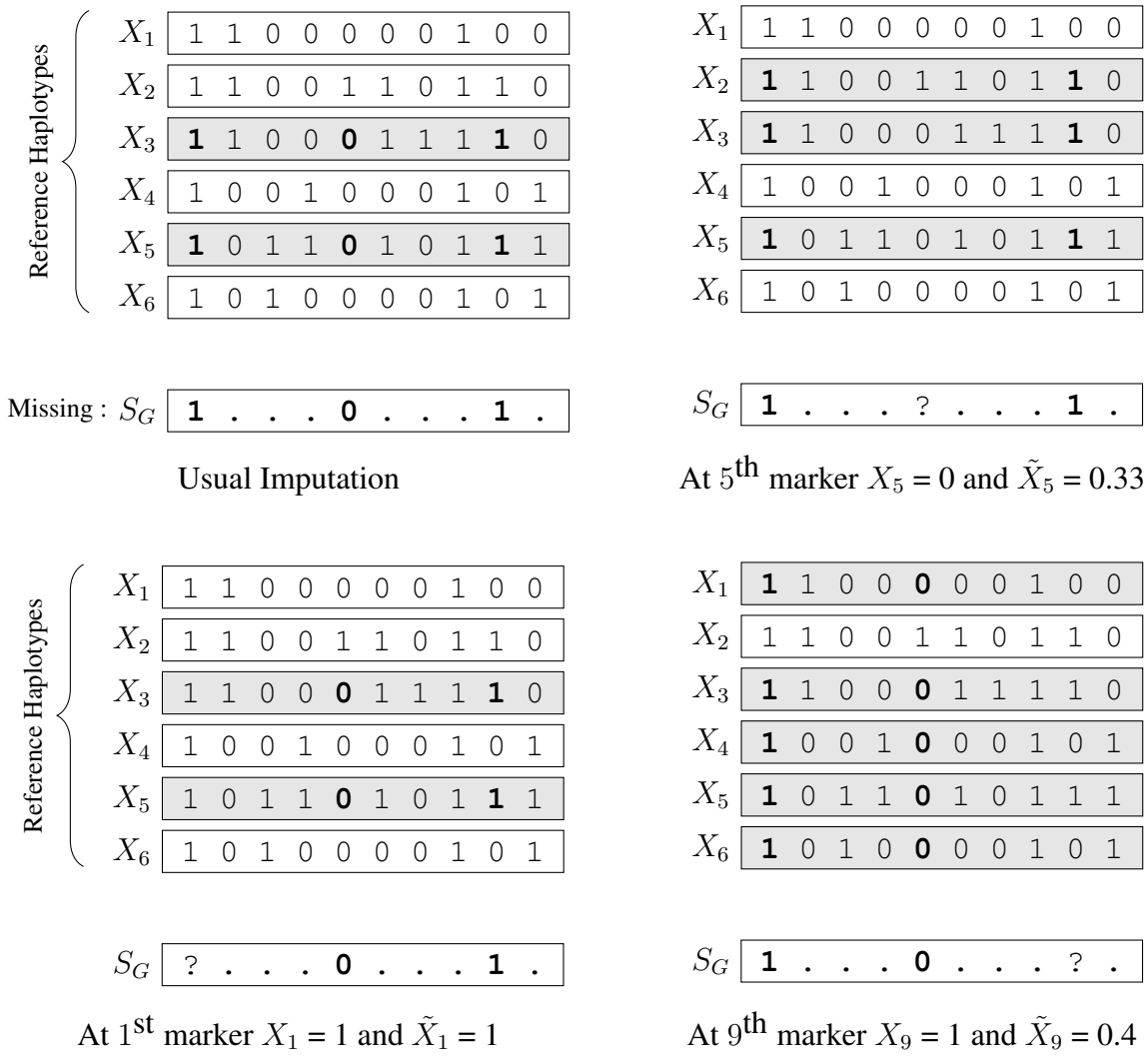


Figure 5.2: Schematic diagram explaining leave-one-out imputation. The figure illustrates leave-one-imputation using a single panel of 6 reference haplotypes ($X_1, \ldots, X_6$) and $M = 10$ markers, where only 3 markers were genotyped ($G = 1, 5, 9$). $S_G$ is the study haplotype with mostly missing. The top left panel shows usual imputation with two haplotype matches ($X_3, X_5$) based on the 3 genotyped positions. The top-right panel shows leave-one-out imputation for marker 5. Masking observed allele in $S_G$ (denoted as '?') we find three haplotype matches ($X_2, X_3, X_5$) based on positions 1 and 9 (marked in bold). This yields a loo imputed dosage of $(1 + 0 + 0)/3 = 0.33$. The bottom-left panel does the same for marker position 1. Hiding that allele on $S_G$ yields 2 reference matches ($X_3$, $X_5$) based on positions 5 and 9 and and a loo dosage of 1.0. In the bottom-right panel, masking observed allele at position 9 similarly yields 6 matches ($X_1$, $X_3, \ldots, X_6$) based on positions 1 and 5 and loo dosage of 0.4.

relative imputation accuracy across the chromosome. In the above example, it seems that the reference panel did comparatively better in the beginning of the genome ($X_1 = 1$ versus $\tilde{X}_1 = 1.0$) than it did towards the end of the genome ($X_9 = 1$ versus $\tilde{X}_9 = 0.4$). The benefit of estimating loo dosages is it provides an ad-hoc way for evaluating the accuracy of imputation based on the genotyped markers only, thereby removing the need for extra genotyping/sequencing for accuracy evaluation.

In minimac3/4, the loo dosages are estimated by setting $P\left(S_k|Y_i\right) = 1$ in equations (3.2) and (3.7) (since hiding the observed allele is equivalent to using a trivial genotype emission probability). Accordingly, the recursive equations for calculation of the loo imputation estimates at genotype markers are as follows:

$$\tilde{\mathfrak{L}}_B\left(Z_i\right) = \left[\left[1 - \lambda_{A:B}\right]\tilde{\mathfrak{L}}_A\left(Z_i\right) + \frac{M_i\lambda_{A:B}}{H}\sum_{j=1}^{V}\tilde{\mathfrak{L}}_A\left(Z_j\right)\right] \tag{5.2}$$

$$\tilde{\mathfrak{R}}_A\left(Z_i\right) = \frac{M_i\lambda_{A:B}}{H}\left[\sum_{j=1}^{V}\tilde{\mathfrak{R}}_B\left(Z_j\right)\right] + \left[1 - \lambda_{A:B}\right]\tilde{\mathfrak{R}}_B\left(Y_i\right) \tag{5.3}$$

- $A$ and $B$ are two consecutive genotyped markers.
- $\lambda_{A:B}$ denotes the template switch probability between markers $A$ and $B$ (defined in Chapter 4, Section 4.2.1, p.42).
- $Z_1, \ldots, Z_V$ are the states of the aggressively reduced state space (minimac4) and $\tilde{\mathfrak{L}}_i(.)$ and $\tilde{\mathfrak{R}}_i(.)$ are their respective left and right probabilities for leave-one-out estimation (at the $i^{th}$ marker).

The left and right probabilities can be used to yield loo posterior probabilities $\tilde{P}\left(Z_v|\text{GWAS}\right)$ (using Equation (4.8), p.43) which can then be used to obtain loo imputation estimates as follows:

$$\tilde{X}_i = \sum_{v=1}^{V} Z_v \times \tilde{P}\left(Z_v|\text{GWAS}\right) \tag{5.4}$$

### 5.2.3 Estimation of Weights

The weights are estimated by training our model on the leave-one-out imputation estimates. At each of the genotyped markers ($i \in \boldsymbol{G}$), let us denote the genotyped values as $G_i$. We train our model by treating the $G_i$ as our pseudo-meta-imputed dosage, and $\tilde{X}_{k,i}$ as our pseudo-haplotype dosages. We use the Nelder-Mead algorithm (Nelder et al. 1965) to obtain $w_1, \ldots, w_k$ by minimizing the following objective function under the constraints $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$:

$$\Phi(w_1, \ldots, w_k) = \sum_{i \in \boldsymbol{G}} \left[ G_i - \sum_{k=1}^{K} w_k \tilde{X}_{k,i} \right]^2 \tag{5.5}$$

In our internal evaluations, we find that optimizing the above function over genomic chunks ($5 - 10$ mega base positions) to obtain chunk specific estimates of $w_1, \ldots, w_K$ provides better imputation accuracy than optimizing them over the whole chromosome. This might be because different regions of the genome require different weights based on which ancestral parent haplotype that region had been copied from.

## 5.3 Results

### 5.3.1 Meta-imputation in Samples of Admixed Ancestry

Previous studies have shown that admixed samples gained better imputation accuracy from multi-ethnic panels than from those with single ethnic samples (Chanda et al. 2012). We evaluated the benefits of meta-imputing samples of admixed ancestry when relatively homogenous reference panels are used for imputation. We mimicked a typical GWA study using the 1000G WGS samples. The two homogenous reference panels comprised of 600 African (AFR) samples and 500 European (EUR) samples, while the joint (merged) reference panel comprised of 1,103 AFR + EUR samples. We imputed into 61 African-American samples in Southwest USA (ASW) and used aggregate[1]

[1] See footnote on page 20 for definition of aggregate $r^2$

$r^2$ for evaluating imputation accuracy. The results show that the AFR panel yields much higher accuracy compared to the EUR panel (see Figure 5.3). For example, the EUR panel is unable to impute any markers with MAF $< 1\%$ while the AFR panel yields imputation accuracy as high as $r^2 = 35\%$ for variants with MAF around $0.1\%$. However, the joint reference panel brings a significant improvement over both panels by increasing the imputation $r^2$ to $45\%$ at MAF around $0.1\%$.

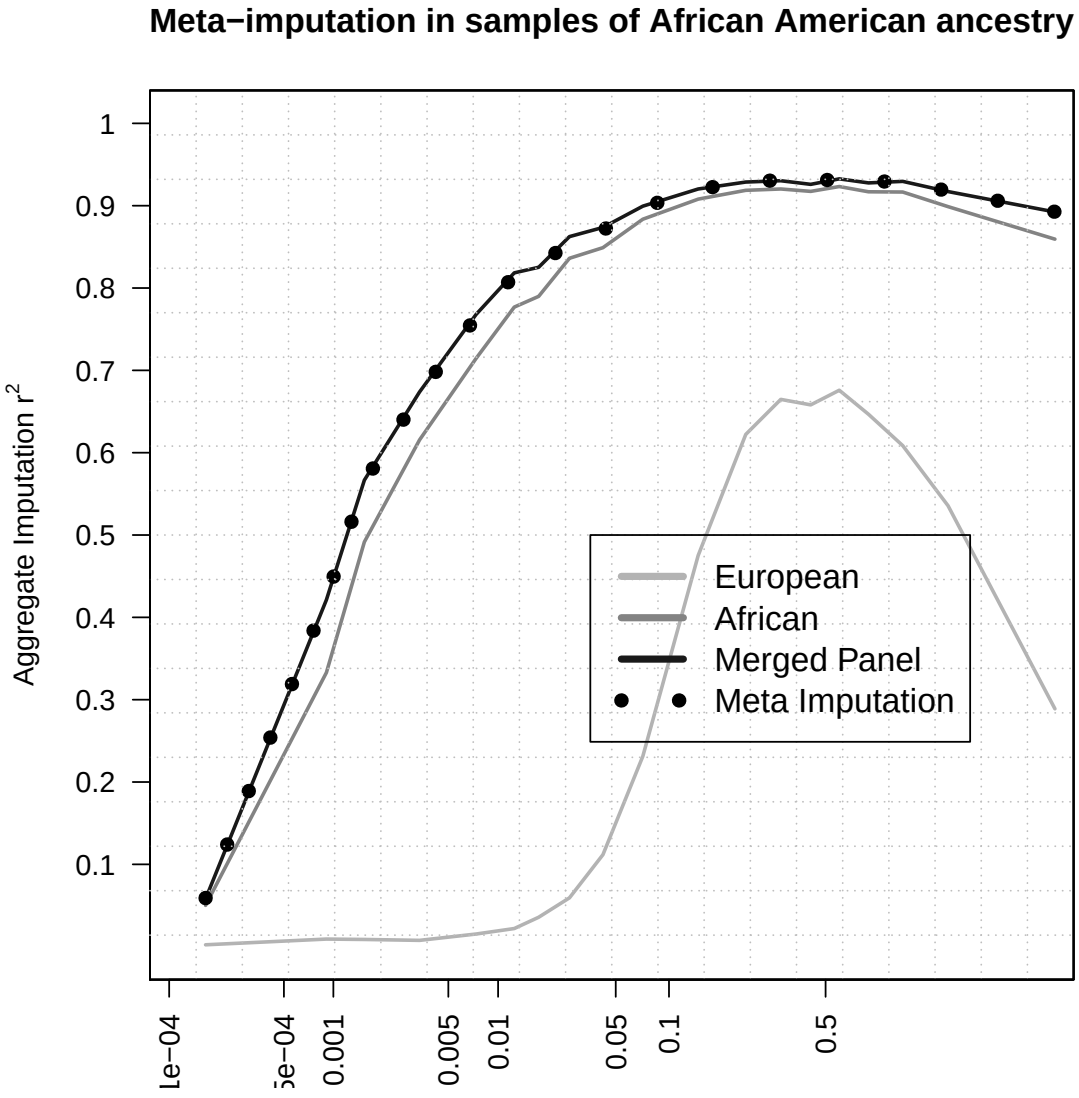


Figure 5.3: Meta-imputation in African American samples. The figure illustrates the imputation accuracy of different panels and methods in samples of African American Ancestry. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$.

Our results also show that meta-imputation of individual imputed dosages yields as much accuracy, if not more, as the joint reference panel (detailed values are given in Table G.1. p.99 in Appendix G). For most part of the rarer frequency spectrum, joint imputation does slightly better than meta-imputation (maximum improvement in imputation $r^2$ is $0.0033$). Although, for the lowest MAF bin [~$10^{-3}$], the meta-imputation $r^2$ is $42.3\%$ compared to the joint imputation $r^2$ of $42.1\%$. However, meta-imputation consistently performs better around the common spectrum, albeit at marginally improved rate (e.g. $r^2$ of $92.8\%$ versus $r^2$ of $92.6\%$ around MAF $\sim 45\%$).

### 5.3.2 Meta-imputation of Three Reference Panels

We next evaluated the effect on imputation accuracy when meta-imputing three reference panels concurrently. Our pseudo-GWA study comprised of 1,000 samples of both pure and admixed ancestry from the TOPMed WGS dataset. Our three reference panels are 661 AFR samples, 503 EUR samples, and 504 East Asian (EAS) samples (from 1000G Phase 3). The merged reference panel is comprised of 1,668 AFR+EUR+EAS samples . The results are shown in Figure G.1, p.97 in Appendix G. The poor performance of the EAS panel could be because less than 10% of the pseudo-GWA study samples are Asian or East Asian (see Figure F.1, p.95 in Appendix F). The results in Figure G.1 show that meta-imputation does almost as well as the joint imputation even when merging data imputed against 3 reference panels. The maximum difference in aggregate $r^2$ is ~$0.016$ and was observed around alternate AF bin $[0.3\% - 0.7\%]$ (detailed values in Table G.1. p.99 in Appendix G).

### 5.3.3 Additional Benefit in Saving Computational time

Here we illustrate an additional benefit of meta-imputation in saving computation time. Large reference panels like HRC or TOPMed can be randomly split into multiple non-overlapping parts, imputed against each such part, then meta-impute the individual imputed dosages. The individual imputations can be run in parallel and thus save in net computation time. Figure G.2, p.98 in Appendix G shows that we lose negligible imputation accuracy by meta-imputation of three disjoint

pieces of the HRC panel compared to the full HRC panel. Apart from the lowest alternate AF bin, meta-imputation does slightly better than the joint imputation. The maximum improvement in aggregate $r^2$ is ~$0.005$ and was observed around alternate AF bin $[0.3\% - 0.7\%]$ (detailed values in Table G.1. p.99 in Appendix G). Additionally, we saved ~35% of original run-time in this study of meta-imputation of three split panels.

### 5.4 Discussion

In this study we aim to provide a coherent and convenient framework for meta-imputation of GWA studies that have been imputed against different reference panels. We showed that meta-imputation compares well to imputation in terms of accuracy. Modern developments in next generation sequencing have now made it feasible to obtain thousands of genomes in a typical genetic study. However, until whole genome sequencing is practical for even larger GWA studies, researchers must rely on imputation to have enough power to efficiently investigate rare mutations potentially associated with a disease. While multiple sequencing projects are providing an ever-increasing abundance of information, a meticulous consolidation of this genetic information is still a challenging problem, both due to enormous computational requirements and to other data sharing restrictions. Meta-imputation resolves most of these issues and enables researchers to merge their imputed datasets in a data-driven dynamic way that attains similar, if not better, accuracy as the consolidated reference panel.

The main caveat with our methodology is that variants which are present in only one reference panel gain nothing from the other panels. Special scenarios where the two reference panels have a small number of overlapping sites might benefit more from alternative methods like "joint imputation" or "cross-imputation". An interesting observation from this study was the slightly superior performance of meta-imputation compared to joint imputation near the rare MAF spectrum. The reason behind this might be, if one of the split panels end up with all copies of the rare allele, the allele frequency increases for that variant in that panel, which consequently improves accuracy. For instance, suppose we split the reference panel into 3 equal parts, and one of the split parts ends

up with all copies of the rare allele while the remaining two have none. The proportion of rare alleles becomes triple in one of the split panels and this increases the imputation accuracy at that variant due to that panel (compared to the merged panel). Since the remaining two panels have no variation at the that site, they are unable to impute these sites at all. The meta-imputation algorithm dynamically estimates the relative efficacy of each split panel and gives all of its weight to the first split panel. Thus, the imputation estimates for this rare variant from meta-imputation might be of higher quality that those from joint imputation. We also note that for common variants, the chances of all copies of the rare allele ending up in a single split panel is comparatively much lower. For instance, if we split a large panel like the HRC into three equal parts, the chances of all copies ending up in one split part is $33\%$ for a variant with MAC $= 2$ and $< 0.01\%$ for a variant with MAC $= 10$. This explains the reason why the improvement is noticeable along the rarer variants but not for common variants.

# CHAPTER 6

## Conclusion

We accomplished the following aims in this dissertation (i) create a reference panel that increases imputation accuracy for European samples, (ii) develop tools for significantly faster genotype imputation with negligible falls in accuracy, (iii) provide web-based solutions to override issues of data privacy, and (iv) establish methods of consolidating multiple reference panels. Our efforts have been to build tools/platforms that aid genetic researchers in next generation of human gene studies. The HRC is currently the most used reference panel on the Michigan imputation server accounting for more than 60% of the total number of genomes imputed. The TOPMed panel is awaiting data permissions before it can be hosted as a panel. The Michigan server has close to ~2,000 users to date and has imputed over ~7 million whole genomes. This server is used by investigators from across the world, but mainly from the US and Europe (Figure I.1, p.105). Our server has also been on social media with users tweeting about their experiences (Figure I.2, p.105). Minimac3 is the current imputation engine behind the Michigan server. We are running more tests on minimac4 before we implement it on the server.

The imputation tools described in this dissertation, minimac3 and minimac4, are the next installments in the series of tools that started with MaCH (Li et al. 2010). MaCH performed imputation and phasing simultaneously. The idea of pre-phasing led to the development of minimac (Howie et al. 2012) and further software engineering techniques led to minimac2 (Fuchsberger et al. 2014). We implemented the idea of state space reduction and its aggressive version to develop minimac3 (Das et al. 2016) and minimac4 respectively. The MaCH/minimac tools are summarized in Figure 6.1. When imputing against the TOPMed reference panel, minimac4 is ~12,000 times

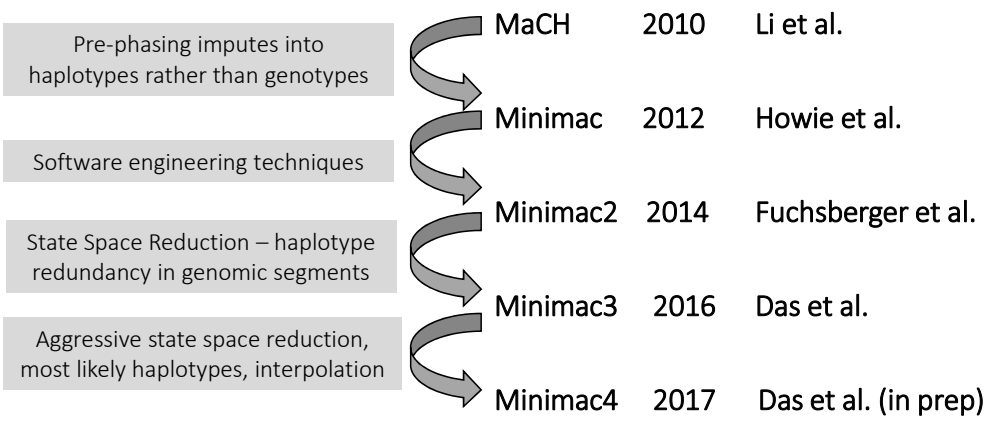


Figure 6.1: Summary of development from MaCH (2010) to Minimac4 (2017).

faster than minimac, while minimac is over 20,000 times faster than MaCH. Assuming one is using a commercial cloud computing system (e.g. Amazon® Web Services at a rate of ~3-4 cents per CPU hour), it would take a researcher using minimac4 only $9 to impute a small genetic study of 1,000 GWA samples, whereas the same would have cost over $20 million back in 2010 (see Figure 6.2). Another way to consider this result is, if we fixed the computation time, at say ~17.5 CPU-minutes per genome (the current time to impute against the TOPMed panel with ~20,000 samples), we could only impute against ~10 samples in 2010. This would have substantially decreased our

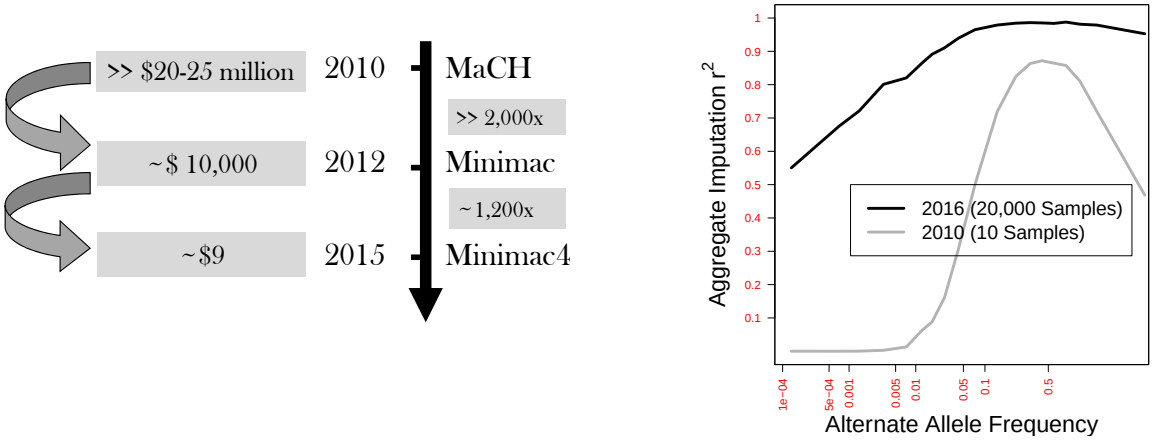


Figure 6.2: Cost savings in imputation and improvement since 2010. The left figure illustrates the savings in money since 2010 if using Amazon® Web Services at a rate of ~3-4 cents per CPU hour. The right figure illustrates the improvement in imputation accuracy at a fixed computational time of ~17.5 minutes per genome, the current time to impute against the TOPMed panel.

imputation accuracy (see Figure 6.2). A take away message from this dissertation is: if we can't build better methods, we should build faster methods, and the amount of data available might make it a better method.

# APPENDIX A

## A brief review of hidden Markov models

### A.1 A Brief Review

A hidden Markov model HMM is a statistical process that models a Markov chain with unobserved (hidden) states. It is a Markov chain where the original states of the process are not observed, but another output, that is dependent on the original but unobserved state through a predetermined probability function, is observed. Figure A.1 briefly explains the working of a HMM. Although the sequence of the original states of the Markov chain are not observed, the sequence of observed tokens generated by the hidden sates gives us some information about the underlying states of the process and hence can be used to make inferences.

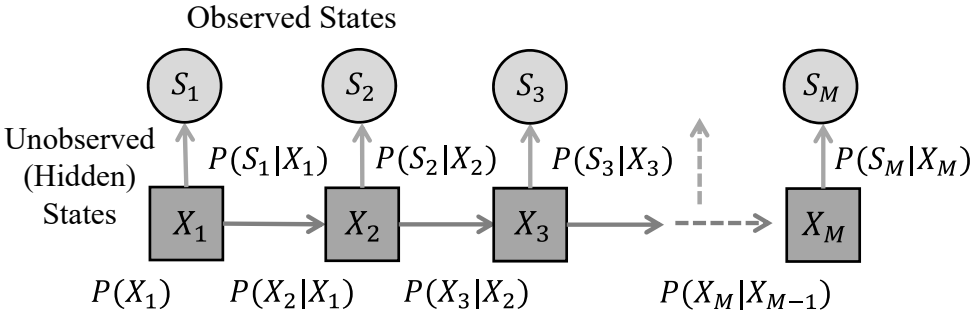


Figure A.1: Illustration of hidden Markov model. The dark gray boxes represent the underlying Markov chain with initial probability $P(X_1)$ and transition functions $P(X_i|X_{i-1})$. The transition function follows the same conditions as a usual Markov chain i.e. $P(X_{i+1}|X_i, X_{i-1}, \ldots, X_1) = P(X_{i+1}|X_i)$. The variables $X_1, \ldots, X_M$ are not observed. The light gray circles represent the observed variables $S_i$ that depend on the unobserved variables $X_i$ through the emission probability function $P(S_i|X_i)$.

HMMs have been used in modelling time series data, speech recognition systems, computa-

tional molecular biology, data compression, pattern recognition, and other areas of artificial intelligence (Ghahramani 2001). It has been used as a standard model for genotype imputation as it suitably fits the following framework:

- '$i$' denotes each marker. The hidden states $X_i$ denote the underlying haplotype at each marker $i$ while the observed state $S_i$ denotes the observed allele (if the observed site was genotyped) or is missing.
- $P(S_i|X_i)$ denotes the genotype emission probability and is equal to $\epsilon$ (genotyping error) if the underlying and observed alleles match at the marker, and is equal to $1-\epsilon$ if they don't match . If the marker was not genotyped then $P(S_i|X_i) = 1$.
- $P(X_{i+1}|X_i)$ denotes the transition function and measures the probability of a template switch between consecutive markers (as in Figure 1.2, p.8). It depends on the recombination rate between markers ($\lambda_i$) and is defined as follows:

$$P(X_{i+1}|X_i) = \begin{cases} 1 - \lambda_i + \frac{\lambda_i}{H} & \text{if } X_{i+1} = X_i \\ \\ \frac{\lambda_i}{H} & \text{if } X_{i+1} \neq X_i \end{cases}$$

The intuition behind the above formulas have been explained in Li et al. 2010. HMMs are one of the most common statistical models employed in genotype imputation. They can be suitably used to infer the underlying haplotype or yield set of probabilities for each possible haplotype at every marker. The mathematics behind the HMM were largely developed by L. E. Baum (Baum et al. 1966), including the famous Baum-Welch's forward and backward equations (Baum et al. 1970) that greatly reduces the algorithmic complexity to compute posterior probability of the hidden states conditional on the observed states i.e. $P(X_i|S_1, \ldots, S_M)$.

## A.2 A Review of HMM in Minimac

Below we give a brief review of the formulas used in minimac. Consider a reference with $H$ haplotypes and a genomic segment bounded by markers $P$ and $Q$. Label the original haplotypes as $X_1, X_2, \ldots, X_H$ (the state space).

### Forward Equations

Let $L_k(.)$ denote the left probabilities for each state at marker '$k$' (for a description of left and right probabilities see Li et al. 2010). $P\left(S_{k+1}|X_l\right)$ denotes the genotype emission probability and is defined above. Below is Baum-Welch's forward equation:

$$L_{k+1}\left(X_l\right) = \left[\left(1-\lambda_k\right) L_k\left(X_l\right) + \frac{\lambda_k}{H}\sum_{j=1}^{H} L_k\left(X_j\right)\right] \times P\left(S_{k+1}|X_l\right) \tag{A.1}$$

**Backward Equations** Similarly, if $R_k(.)$ denote the right probabilities for each state at marker '$k$', the following equation gives the Baum-Welch's backward equations.

$$R_{k-1}\left(X_l\right) = \frac{\lambda_{k-1}}{H}\left[\sum_{j=1}^{H} R_k\left(X_j\right) P\left(S_k|X_j\right)\right] + \left(1-\lambda_{k-1}\right) R_k\left(X_l\right) P\left(S_k|X_l\right) \tag{A.2}$$

**Final Imputation Formula** The final formula gives the posterior probability of state.

$$P\left(X_l\,|\text{GWAS}\right) \;=\; L_K(X_l) \times R_K(X_l) \tag{A.3}$$

MaCH uses similar formulas but its state space comprises of all possible genotype configurations possible from $H$ haplotypes (complexity $O(H^2)$), while minimac uses all possible haplotypes as its state space (complexity $O(H)$). These formulas provide a base for developing models for minimac3/4.

# APPENDIX B

## Asymptotic results for association test with imputed genotypes

In a GWA study, for variants not directly assayed on the chip, investigators need to use imputed dosages for testing association with a physical trait. In this appendix we will look at some asymptotic results of the effect size estimates from imputed dosages, including any bias compared to using unobserved genotypes. For sake of simplicity, we assume everything at a haplotype level i.e. for a fixed marker the observed allele at a haplotype is denoted as $G_i$ (where $G_i = 0, 1$), and the imputed dosage is denoted as $D_i$ (where $0 \leq D_i \leq 1$) for $i = 1, \ldots, 2n$ (where $n$ is the number of samples). In the sections below, we first show that in a simple linear regression the estimated effects sizes based on the imputed dosage are consistent for the true effect sizes of the variant, provided the imputed variant is "well-calibrated". An imputed variant is said to be well-calibrated for a sample if the probability of observing the '1' allele for that sample is equal to its imputed dosage, that is $E(G|D) = D$. Intuitively speaking, a variant would be well-calibrated if one or more template matches were found around that variant, and would be poorly calibrated if no matches were found. We also show that for the difference in proportion test, using the imputed variant yields an effective sample size of $\rho^2 n$ where $\rho$ is the correlation between the genotype and the dosage.

### B.1 Simple Linear Regression

Let us assume a linear regression model where the observed alleles are the independent variables ($G_i$, treated as continuous) and the phenotype is the outcome variable ($Y_i$).

$$Y_i = \alpha + \beta G_i + \epsilon_i$$

Let us denote the usual least-square estimator based on the unobserved genotypes/alleles as $\hat{\beta}_{\text{genotyped}}$. Instead of observing the alleles, we have imputed dosages ($D_i$) which are to be used as proxies for the unobserved alleles. Accordingly, our least-square estimator based on imputed dosages would be

$$\hat{\beta}_{\text{imputed}} = \frac{S_{DY}}{S_{DD}}$$

where $S_{DY} = \sum D_i Y_i - 2n\bar{D}\bar{Y}$ is the sample covariance between genotyped and dosages and $S_{DD} = \sum (D_i - \bar{D})^2$ is the sample variance of the dosages. We first simplify the above expression by substituting $Y_i = \alpha + \beta G_i + \epsilon_i$ as follows:

$$\begin{aligned}
\hat{\beta}_{\text{imputed}} &= \frac{\sum_{i=1}^{2n} D_i Y_i - 2n\bar{D}\bar{Y}}{S_{DD}} \\
&= \frac{\alpha \sum D_i + \beta \sum D_i G_i + \sum D_i \epsilon_i - 2n\alpha\bar{D} + 2\beta n\bar{G} + 2n\bar{D}\bar{\epsilon}}{S_{DD}} \\
&= \frac{\beta S_{DG} + \sum D_i \epsilon_i + 2n\bar{D}\bar{\epsilon}}{S_{DD}}
\end{aligned}$$

where $S_{DG} = \sum D_i G_i - 2n\bar{D}\bar{G}$. Since the imputed dosages can be assumed to be independent of the errors and $E(\epsilon) = 0$, we have:

$$E(\hat{\beta}_{\text{imputed}}) = \beta E\left(\frac{S_{DG}}{S_{DD}}\right)$$

Assuming sufficiently large sample we can assume $\frac{S_{DG}}{S_{DD}} \approx \rho\frac{\sigma_G}{\sigma_D}$ (by law of large numbers). Thus, $\hat{\beta}_{imputed}$ is an consistent estimator for $\beta^* = \beta\left[\rho\frac{\sigma_G}{\sigma_D}\right]$. We next show that under the assumption of 'well-calibration' $\beta^*$ is equal to $\beta$. This follows from the following set of equations:

$$E(G) = E(E(G|D)) = E(D)$$

$$E(DG) = E(E(DG|D)) = E(DE(G|D)) = E(D^2)$$

$$\text{Cov}(D, G) = E(DG) - E(D)E(G) = V(D)$$

$$\rho = \text{Corr}(D, G) = \frac{\text{Cov}(D, G)}{\sqrt{V(G)V(D)}} = \sqrt{\frac{V(D)}{V(G)}}$$

Thus, under the assumption of well-calibration we have $\rho = \frac{\sigma_D}{\sigma_G}$ which implies that $\beta^* = \beta$. Hence, in a linear regression, given a large sample, the estimated effect size from the imputed data is unbiased for the population effect size. We next compare the standard errors of the statistics.

$$\text{Var}\left(\hat{\beta}_{\text{imputed}}\right) = \frac{\sigma^2}{S_{DD}} \approx \frac{\sigma^2}{\sigma_D} = \frac{\sigma^2}{\rho\sigma_G} \approx \frac{\text{Var}\left(\hat{\beta}_{\text{genotyped}}\right)}{\rho}$$

The above results show that $\hat{\beta}_{\text{imputed}}$ has a larger standard error than $\hat{\beta}_{\text{genotyped}}$ and the power for testing association using imputed data yields same power as using genotypes with $\rho^2 n$ sample size.

### B.2 Difference in Proportion Test

This test analyzes the difference in the allele frequency between cases and controls. Let $p_{\text{Case}}$ (and $p_{\text{Control}}$) denote the population allele frequencies in cases (and controls) i.e. $p_{\text{Case}} = P(G = 1|\text{ Case})$ and $p_{\text{Control}} = P(G = 1|\text{ Control})$. The test statistic for testing $H_0 : p_{\text{Case}} = p_{\text{Control}}$ is given as follows (for a setup with $n_1$ cases and $n_2$ controls):

$$T_{\text{genotyped}} = \frac{\hat{p}_{\text{Case}} - \hat{p}_{\text{Control}}}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

When we use imputed dosages as our proxies for the alleles the above statistic becomes:

$$T_{\text{imputed}} = \frac{\bar{D}_{\text{Case}} - \bar{D}_{\text{Control}}}{\sqrt{\bar{D}(1-\bar{D})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Under the assumption of well-calibration we have $E(\bar{D}) = E(D) = E(G) == p$. Thus, for large samples the denominators of $T_{\text{genotyped}}$ and $T_{\text{imputed}}$ are approximately same. We next evaluate

the asymptotic mean of the numerator.

$$\begin{aligned} E\left[\bar{D}_{\text{Case}} - \bar{D}_{\text{Control}}\right] &= E(D|\text{ Case}) - E(D|\text{ Control}) \\ &= E(D|G=1)P(G=1|\text{ Case}) + E(D|G=0)P(G=0|\text{ Case}) \\ &\quad -E(D|G=1)P(G=1|\text{ Control}) - E(D|G=0)P(G=0|\text{ Control}) \\ &= [E(D|G=1) - E(D|G=0)]\left[p_{\text{Case}} - p_{\text{Control}}\right] \end{aligned}$$

We next see that:

$$\begin{aligned} E(D|G=1) - E(D|G=0) &= \int D f_{D|G=1} - \int D f_{D|G=0} \\ &= \frac{\int D f_{D,G}(D,1) - P(G=1)\left[\int D f_D\right]}{P(G=0)P(G=1)} \\ &= \frac{E(D|G=1)P(G=1) - P(G=1)E(D)}{P(G=0)P(G=1)} \\ &= \frac{E(DG) - E(D)E(G)}{V(G)} \\ &= \rho\frac{\sigma_D}{\sigma_G} \\ &= \rho^2 \end{aligned}$$

The last equation follows from $\rho = \frac{\sigma_D}{\sigma_G}$ which was proved earlier as a consequence of well-calibration. Thus, we have:

$$E\left[\bar{D}_{\text{Case}} - \bar{D}_{\text{Control}}\right] = \rho^2\left[p_{\text{Case}} - p_{\text{Control}}\right]$$

Hence, for large samples, the test statistics $T_{\text{genotyped}}$ and $T_{\text{imputed}}$ approximately follow non-central t-distribution with the following means respectively,

$$\left[\frac{p_{\text{Case}} - p_{\text{Control}}}{\sqrt{p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}\right] \quad \text{and} \quad \rho^2\left[\frac{p_{\text{Case}} - p_{\text{Control}}}{\sqrt{p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}\right]$$

Thus, the distribution of $T_{\text{imputed}}$ would approximately be the same as $T_{\text{genotyped}}$ if the actual sample sizes for the cases and controls were $\rho^2 n_1$ and $\rho^2 n_2$ respectively. This proves that if the correlation between the imputed dosage and unobserved alleles is $\rho^2$ then testing using that imputed variant reduces the effective samples size by a factor of $\rho^2$.

## APPENDIX C

### Supplementary information for the Haplotype Reference Consortium

## C.1 Supplementary Tables

| Samples | 1000G Phase 3 (N=2,525) | | | HRC Pilot (N=13,309) | | | HRC Full (N=32,905) | | |
|---|---|---|---|---|---|---|---|---|---|
| Markers | 1000G Phase 3 | HRC MAC5 | HRC MAC5 + marker filters | 1000G Phase 3 | HRC MAC5 | HRC MAC5 + marker filters | 1000G Phase 3 | HRC MAC5 | HRC MAC5 + marker filters |
| REF/REF | 0.10 | 0.10 | 0.10 | 0.07 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 |
| REF/ALT | 0.61 | 0.59 | 0.59 | 0.36 | 0.38 | 0.36 | 0.34 | 0.36 | 0.34 |
| ALT/ALT | 0.43 | 0.40 | 0.40 | 0.27 | 0.26 | 0.25 | 0.25 | 0.24 | 0.23 |
| NRD | 0.70 | 0.67 | 0.67 | 0.43 | 0.43 | 0.42 | 0.41 | 0.41 | 0.40 |

Table C.1: Evaluation of genotype calling process. Discordance rates of genotypes called using different marker lists (Markers row) and sample sets (Samples row). NRD = non reference allele discordance percentage.

| No | Cohort Description | No of Samples | Average Depth | Reference |
|---|---|---|---|---|
| 1 | UK10K Project | 3,715 | 6.5x | (UK10K et al. 2015) |
| 2 | UK Inflammatory Bowel Disease Genetics Consortium | 4,474 | 4x + 2x | (Barrett et al. 2008) |
| 3 | SardiNIA Project | 3,445 | 4x | (Sidore et al. 2015) |
| 4 | Age-related Macular Degeneration | 3,189 | 4x | (Chen et al. 2010) |
| 5 | The Genetics of Type 2 Diabetes Consortium | 2,709 | 4x/Exome | (Fuchsberger et al. 2016) |
| 6 | 1000 Genomes Phase 3 | 2,495 | 4x/Exome | (1000G 2015) |
| 7 | Bipolar Disorder: Improving Diagnosis, Guidance and Education (BRIDGE) | 2,487 | 6-8x (12x) | - |
| 8 | Sequencing Initiative Suomi, Kuusamo | 1,918 | 4x | (Vartiainen et al. 2009) |
| 9 | Minnesota Center for Twin and Family Research | 1,325 | 10x | (Vrieze et al. 2014) |
| 10 | Genetics and Epidemiology of Colorectal Cancer Consortium | 1,130 | 4-6x | (Hays et al. 2003) |
| 11 | The Nord-Trøndelag Health Study | 1,023 | 4x | (Krokstad et al. 2012) |
| 12 | Project MinE | 935 | 45x | - |
| 13 | Genomes of Netherlands | 748 | 12x | (GoNL 2014) |
| 14 | The Genomic Psychiatry Cohort | 697 | 30x | (Pato et al. 2013) |
| 15 | Invecchiare in Chianti Study | 676 | 7x | (Ferrucci et al. 2000) |
| 16 | Nephrotic Syndrome Study Network | 402 | 4x | (Gadegbeku et al. 2013) |
| 17 | Orkney Complex Disease Study | 398 | 4x | (McQuillan et al. 2008) |
| 18 | Italian Network of Genetic Isolates - Fruili Venezia Giulia | 250 | 4-10x | (Colonna et al. 2012) |
| 19 | Hellenic Isolated Cohorts | 247 | 4x (1x) | (Panoutsopoulou et al. 2014) |
| 20 | Val Borbera Isolated Population Project | 225 | 6x | (Esko et al. 2012) |

Table C.2: Summary of studies contributing samples to the HRC panel.

| Number of samples | Studies | | Removal choice |
|---|---|---|---|
| 5 | AMD | AMD | Removed the duplicates randomly. |
| 36 | IBD | IBD | These were duplicates between Crohns and UC studies. Removed the duplicates randomly. |
| 5 | FINLAND | FINLAND | Removed lower coverage (4x) Kuusamo samples in preference to the higher coverage (6x) SiSu samples. |
| 14 | GECCO | GECCO | Removed the duplicates randomly. |
| 17 | GoT2D | FINLAND | Removed the FINLAND samples. |
| 79 | GoT2D | UK10K | Removed the UK10K/UK10K duplicates, then removed randomly otherwise. |
| 34 | GPC | BRIDGES | Removed the GPC samples. |
| 32 | GPC | GPC | Removed the duplicates randomly. |
| 14 | MCTFR | MCTFR | Removed the duplicates randomly. |
| 1 | NEPTUNE | NEPTUNE | Removed the duplicates randomly. |
| 1 | ORCADES | 1000GP3 | Removed the ORCADES sample. |
| 1 | ProjectMinE | GoNL | Removed the ProjectMinE sample. |
| 1 | ProjectMinE | ProjectMinE | Removed the duplicates randomly. |
| 3 | SardinIA | SardinIA | Removed the duplicates randomly. |
| 26 | UK10K | UK10K | These were monozygotic twins already identified by UK10K. Removed based on a list from UK10K of samples they had already excluded from downstream analysis. |

Table C.3: Details of duplicate removal in HRC. Each row of the table details the number of duplicate pairs found within and between studies, together with the method by which duplicates were removed

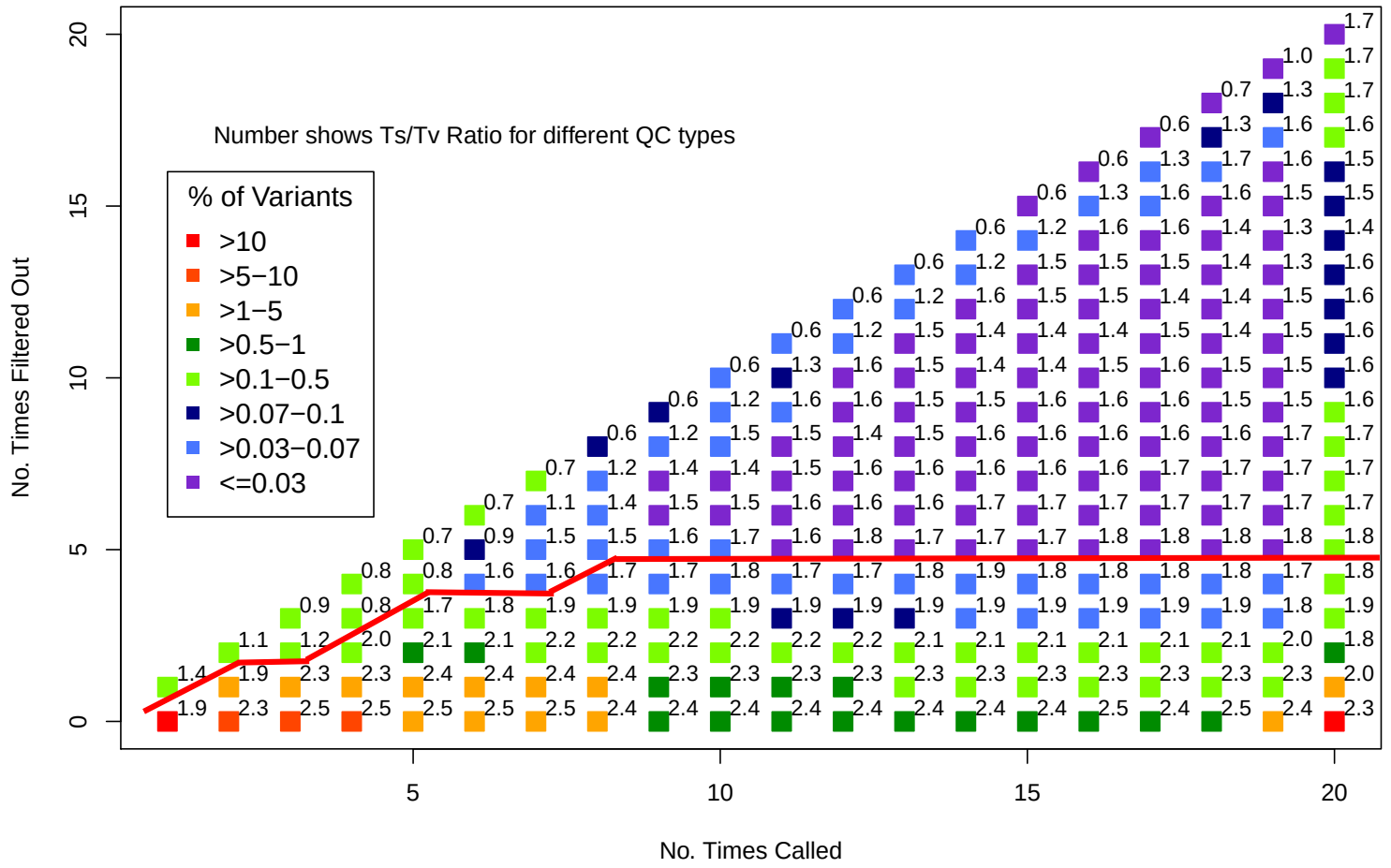


Figure C.1: Illustration of site filtering strategy. This figure summarizes our filtering strategy. On the X-axis we show the number of studies a variant was called in (out of 20) and on the y-axis we show the number of times it was filtered out by the cohort-specific internal QC pipelines. The color shows the percentage of variants in each such cell (red means more than 10% of variants lie in that cell while blue means less than 0.1%). The number to the top right of each cell denotes the Ts/Tv ratio for all sites in that cell. Cells higher in the plot have been filtered out relatively often and usually represent poor variants, as is also seen from the low Ts/Tv ratio. All variants above the red line were filtered out (which excludes all cells which had been filtered independently by more than 4 studies or have Ts/Tv ratio less than 1.7)

# APPENDIX D

## Proof of methods and supplementary information for minimac3

### D.1 Proof of Method Formulas

In this appendix, we will prove that the formulations for the reduced state space HMM used in minimac3 (described in Section 3.2.1) are mathematically equivalent to the original HMM used in minimac (briefly reviewed in Section A.2, Appendix A), and thus would yield the same imputation estimates in the end. To this end, we will first need to prove equation (3.3), p.27 ($\boldsymbol{\mathcal{L}}_Q \rightarrow L_Q$), which essentially states that the left probabilities of the original states can be extracted from the left probabilities of the reduced states (Proposition D.1.1). The proof of equation (3.8), p.28 (which makes the same claim about right probabilities) is very similar and will be skipped in this report. Next, we will prove equation (3.11), p.29 which claims that the posterior probabilities obtained from the reduced states would be numerically same as those obtained from the original state space (Proposition D.1.2), thus finally proving that both the hidden Markov models are mathematically equivalent.

**Proposition D.1.1.** *For any $k$ such that $(P \leq k \leq Q)$ and $X_j$ such that $X_j = Y_i$*

$$L_k(X_j) = \boldsymbol{\mathcal{L}}_k^R(Y_i) \times \left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{L}}_k^{NR}(Y_i) \left[\frac{L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right] \tag{D.1}$$

*Proof.* We will use mathematical induction to prove this claim. Proving it for $k = P$ follows easily from the fact that $\boldsymbol{\mathcal{L}}_P^R = 0$ (obtained by substituting $k = P$ in equations (3.4) and (3.5)).

To prove it for general $k > P$ (under mathematical induction hypothesis), we assume it's true for some $k = K$ and we will prove it here for $k = K + 1$.

To this end, it suffices to prove that expression of $L_{K+1}(X_j)$ from equation (D.1) (for $k = K + 1$) satisfies the actual recursion for the forward equations on the original HMM in minimac (given in equation (A.1), p.74 in Appendix A). To prove that, we first note that equation (3.4) can be re-written in terms of a recursion as follows:

$$\boldsymbol{\mathcal{L}}^{NR}_{K+1}(Y_i) = \boldsymbol{\mathcal{L}}^{NR}_{K}(Y_i) \times [(1 - \lambda_K) P\left(S_{K+1}|Y_i\right)] \tag{D.2}$$

Next, substituting $k = K + 1$ in equation (3.2) we get:

$$\boldsymbol{\mathcal{L}}_{K+1}\left(Y_i\right) = \left[\left[1 - \lambda_K\right] \boldsymbol{\mathcal{L}}_K\left(Y_i\right) + \frac{N_i \lambda_K}{H} \sum_{j=1}^{U} \boldsymbol{\mathcal{L}}_K\left(Y_j\right)\right] \times P\left(S_{K+1}|Y_i\right) \tag{D.3}$$

Substituting the expression for $\boldsymbol{\mathcal{L}}_{K+1}$ from the above equation (D.3) in equation (3.5) (for $k = K + 1$), we get the following:

$$\boldsymbol{\mathcal{L}}^{R}_{K+1}\left(Y_i\right) = \left[\left[1 - \lambda_K\right] \boldsymbol{\mathcal{L}}^{R}_K\left(Y_i\right) + \frac{N_i \lambda_K}{H} \sum_{j=1}^{U} \boldsymbol{\mathcal{L}}_K\left(Y_j\right)\right] \times P\left(S_{K+1}|Y_i\right) \tag{D.4}$$

Finally, substituting the values of $\boldsymbol{\mathcal{L}}^{NR}_{K+1}\left(Y_i\right)$ and $\boldsymbol{\mathcal{L}}^{R}_{K+1}\left(Y_i\right)$ from equations (D.2) and (D.4) into the right hand side of equation (D.1) (for $k = K + 1$), we get:

$$\begin{aligned}
& RHS \\
= \;& \boldsymbol{\mathcal{L}}^{R}_{K+1}(Y_i) \times \left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{L}}^{NR}_{K+1}(Y_i) \left[\frac{L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right] \\
= \;& \left[\frac{\left(1 - \lambda_K\right) \boldsymbol{\mathcal{L}}^{R}_K\left(Y_i\right)}{N_i} + \frac{\lambda_K \sum_{j=1}^{U} \boldsymbol{\mathcal{L}}_K\left(Y_j\right)}{H} + \frac{\boldsymbol{\mathcal{L}}^{NR}_K(Y_i)\left(1 - \lambda_K\right) L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right] P\left(S_{K+1}|Y_i\right) \\
= \;& \left[\left(1 - \lambda_K\right)\left[\boldsymbol{\mathcal{L}}^{R}_K\left(Y_i\right)\left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{L}}^{NR}_K(Y_i)\left[\frac{L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right]\right] + \frac{\lambda_K \sum_{j=1}^{U} \boldsymbol{\mathcal{L}}_K\left(Y_j\right)}{H}\right] P\left(S_{K+1}|Y_i\right) \\
= \;& \left[\left(1 - \lambda_K\right)\left[L_K\left(X_j\right)\right] + \frac{\lambda_K \sum_{i=1}^{H} L_K\left(X_i\right)}{H}\right] \times P\left(S_{K+1}|Y_i\right)
\end{aligned}$$

$$= L_{K+1}(X_j)$$
$$= LHS$$

The last step follows from equation (A.1), p.74 and the step before that follows from the induction hypothesis that the proposition is true for $k = K$ (i.e. $L_K(X_j) = \boldsymbol{\mathcal{L}}_K^R(Y_i)\left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{L}}_K^{NR}(Y_i)\left[\frac{L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right]$), and from the identity $\sum_{i=1}^{H} L_K(X_i) = \sum_{j=1}^{U} \boldsymbol{\mathcal{L}}_K(Y_j)$ (which follows from Equation (3.1)). □

**Proposition D.1.2.** *The posterior probability of each reduced state $Y_i$ is given as:*

$$\begin{aligned} P\left(Y_i \,|\mathit{GWAS}\right) &= \left[\sum_{j=1}^{H} I(X_j = Y_i) L_P\left(X_j\right) R_Q\left(X_j\right)\right] \times \left[\frac{\boldsymbol{\mathcal{L}}_K\left(Y_i\right)}{\boldsymbol{\mathcal{L}}_P\left(Y_i\right)} \times \frac{\boldsymbol{\mathcal{R}}_K\left(Y_i\right)}{\boldsymbol{\mathcal{R}}_Q\left(Y_i\right)}\right] \\ &\quad + \frac{1}{N_i}\left[\boldsymbol{\mathcal{L}}_K\left(Y_i\right)\boldsymbol{\mathcal{R}}_K\left(Y_i\right) - \boldsymbol{\mathcal{L}}_K^{NR}\left(Y_i\right)\boldsymbol{\mathcal{R}}_K^{NR}\left(Y_i\right)\right] \end{aligned} \tag{D.5}$$

*Proof.* To prove this, we start from the LHS of the above equation (and using equation (A.3), p.74 in Appendix A):

$$\begin{aligned} & P\left(Y_i \,|\text{GWAS}\right) \\ = & \sum_{j=1}^{H} I(X_j = Y_i) P\left(X_j|\text{GWAS}\right) \\ = & \sum_{j=1}^{H} I(X_j = Y_i) L_K(X_j) R_K(X_j) \\ = & \sum_{j=1}^{H} I(X_j = Y_i)\left(\boldsymbol{\mathcal{L}}_k^R(Y_i) \times \left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{L}}_k^{NR}(Y_i)\left[\frac{L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right]\right) \\ & \qquad \times \left(\boldsymbol{\mathcal{R}}_k^R(Y_i) \times \left[\frac{1}{N_i}\right] + \boldsymbol{\mathcal{R}}_k^{NR}(Y_i)\left[\frac{R_Q(X_j)}{\boldsymbol{\mathcal{R}}_Q(Y_i)}\right]\right) \\ = & \left[\sum_{j=1}^{H} I(X_j = Y_i) L_P(X_j) R_Q(X_j)\right] \times \left[\frac{\boldsymbol{\mathcal{L}}_K^{NR}(Y_i)}{\boldsymbol{\mathcal{L}}_P(Y_i)} \times \frac{\boldsymbol{\mathcal{R}}_K^{NR}(Y_i)}{\boldsymbol{\mathcal{R}}_Q(Y_i)}\right] \\ & \qquad + \left(\frac{\boldsymbol{\mathcal{L}}_K^{NR}(Y_i)\boldsymbol{\mathcal{R}}_K^R(Y_i)}{N_i}\right) \times \left(\frac{\sum_{j=1}^{H} I(X_j = Y_i) L_P(X_j)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\right) \end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{\boldsymbol{\mathcal{L}}_K^R(Y_i)\boldsymbol{\mathcal{R}}_K^{NR}(Y_i)}{N_i}\right)\times\left(\frac{\sum_{j=1}^H I(X_j=Y_i)R_Q(X_j)}{\boldsymbol{\mathcal{R}}_Q(Y_i)}\right)\\
&+\frac{\sum_{j=1}^H I(X_j=Y_i)\boldsymbol{\mathcal{L}}_K^R(Y_i)\boldsymbol{\mathcal{R}}_K^R(Y_i)}{N_i^2}\\
=&\left[\sum_{j=1}^H I(X_j=Y_i)L_P(X_j)R_Q(X_j)\right]\times\left[\frac{\boldsymbol{\mathcal{L}}_K^{NR}(Y_i)}{\boldsymbol{\mathcal{L}}_P(Y_i)}\times\frac{\boldsymbol{\mathcal{R}}_K^{NR}(Y_i)}{\boldsymbol{\mathcal{R}}_Q(Y_i)}\right]\\
&+\frac{1}{N_i}\left[\boldsymbol{\mathcal{L}}_K^R(Y_i)\boldsymbol{\mathcal{R}}_K^R(Y_i)+\boldsymbol{\mathcal{L}}_K^{NR}(Y_i)\boldsymbol{\mathcal{R}}_J^R(Y_i)+\boldsymbol{\mathcal{L}}_K^R(Y_i)\boldsymbol{\mathcal{R}}_K^{NR}(Y_i)\right]\\
=&\left[\sum_{j=1}^H I(X_j=Y_i)L_P\left(X_j\right)R_Q\left(X_j\right)\right]\times\left[\frac{\boldsymbol{\mathcal{L}}_K\left(Y_i\right)}{\boldsymbol{\mathcal{L}}_P\left(Y_i\right)}\times\frac{\boldsymbol{\mathcal{R}}_K\left(Y_i\right)}{\boldsymbol{\mathcal{R}}_Q\left(Y_i\right)}\right]\\
&+\frac{1}{N_i}\left[\boldsymbol{\mathcal{L}}_K\left(Y_i\right)\boldsymbol{\mathcal{R}}_K\left(Y_i\right)-\boldsymbol{\mathcal{L}}_K^{NR}\left(Y_i\right)\boldsymbol{\mathcal{R}}_K^{NR}\left(Y_i\right)\right]
\end{aligned}
$$

This proves the second proposition. □

### D.2 Supplementary Tables

| 1000G Super Population | No. of segments in whole genome | Average No of variants per segment | Average No (S.D.) of unique haplotypes per segment | Average time (in seconds) to impute single sample | Average time (in seconds) to impute whole genome |
|---|---|---|---|---|---|
| African Americans | 49584 | 36 | 14 (2.6) | 11.65 | 582.5 |
| Hispanics | 43522 | 41 | 13 (2.5) | 10.83 | 541.5 |
| South Asians | 40031 | 44 | 12 (2.4) | 10.42 | 521.0 |
| East Asians | 35822 | 49 | 10 (2.3) | 9.91 | 495.5 |
| Europeans | 35367 | 50 | 10 (2.4) | 9.74 | 487.0 |

Table D.1: Summary of data compression for different ancestries. The ancestries are the different super populations of 1000G Phase 3

| Reference panel | No. of samples | No. of markers | Unzipped format [GB] | | | Zipped format [GB] | | |
|---|---|---|---|---|---|---|---|---|
| | | | VCF | m3vcf | % saving | VCF | m3vcf | % saving |
| 1000G Phase 1 | 1,092 | 617,694 | 2.6 | 0.17 | 93.65 | 0.08 | 0.03 | 61.9 |
| Sardinia | 3,489 | 331,760 | 1.9 | 0.1 | 94.58 | 0.07 | 0.02 | 70.83 |
| 1000G Phase 3 | 2,004 | 1,047,613 | 9.9 | 0.34 | 96.56 | 0.22 | 0.06 | 70.73 |
| HRC | 32,390 | 885,404 | 108 | 1.8 | 98.33 | 1.8 | 0.28 | 84.44 |

Table D.2: Comparison of the VCF and the m3vcf file format for different reference panels (1000G Phase 1, 1000G Phase 3, Sardinia and HRC).

# APPENDIX E

## Proof of methods and supplementary diagrams for minimac4

### E.1 Proof of Method Formulas

In minimac4, we implement an aggressive version of the state space reduction method that was applied in minimac3. In this section we prove that the modified formulas of Baum-Welch equation (as described for minimac4 in Section 4.2.1, p.39) implemented on the genotyped markers using the modified version of of recombination rate (see Eq. (4.1), p.42), yields the same posterior probabilities as in the original state space. Additionally, we note that the aggressive reduced states, although might disagree at ungenotyped markers, should yield the same mathematically equivalent HMM model as the original model when applied on all markers across the chromosome instead of only genotyped markers. The only caveat would be that we cannot generate allelic probabilities from the posterior probabilities of aggressively reduced states (like we could in minimac3). This is because in minimac3, haplotypes collapsed in the same reduced state agree at all markers. Thus, each reduced state in minimac3 had the same allele. In minimac4, haplotypes collapsed together in the same state might have different alleles at ungenotyped markers. There is no easy formula to split the posterior probability into different components based on which allele is observed at the marker. In this section, we first prove Eq. (4.3), p.42 which shows that the Baum-Welch's forward equation can be modified to implement only on the genotyped markers to yield the same mathematical model as in minimac3. In the following proportions, $A$ and $B$ be two consecutive genotyped markers.

**Proposition E.1.1.** *For $K$ such that $A < K < B$, we have $\sum_{i=1}^{V} \mathfrak{L}_K(Z_i) = \sum_{i=1}^{V} \mathfrak{L}_A(Z_i)$*

*Proof.* We first prove this for $K = A+1$. It will follow easily for $K > A+1$. The proof hinges on the fact that all sites in between $A$ and $B$ are ungenotyped and thus have the emission probability $P(S_i|Z) = 1$ for all $Z \in \boldsymbol{\mathcal{S}}_3$ and $A < i < B$. Substituting $k = A+1$ in Equation (3.2) on page 27 (note that for the space $\boldsymbol{\mathcal{S}}_3$ we have to replace $\boldsymbol{\mathcal{L}}(.)$ with $\boldsymbol{\mathfrak{L}}(.)$ and $N_i$ with $M_i$):

$$\begin{aligned}
& \sum_{i=1}^{V} \boldsymbol{\mathfrak{L}}_{A+1}(Z_i) \\
= & \sum_{i=1}^{V} \left[ (1-\lambda_A)\, \boldsymbol{\mathfrak{L}}_A(Z_i) + \frac{M_i \lambda_A}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_A(Z_j) \right] \times P(S_{A+1}|Z_i) \\
= & (1-\lambda_A) \sum_{i=1}^{V} \boldsymbol{\mathfrak{L}}_A(Z_i) + \frac{\lambda_A \sum_{i=1}^{V} M_i}{H} \left[ \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_A(Z_j) \right] \\
= & (1-\lambda_A) \sum_{i=1}^{V} \boldsymbol{\mathfrak{L}}_A(Z_i) + \lambda_A \left[ \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_A(Z_j) \right] \\
= & \sum_{i=1}^{V} \boldsymbol{\mathfrak{L}}_A(Z_i)
\end{aligned}$$

□

The next proposition proves that it suffices to implement the HMM on the genotype markers only using the modified version recombination rate.

**Proposition E.1.2.** *If $\boldsymbol{\mathfrak{L}}_B(.)$ is generated by applying Equation (3.2) on page 27 recursively on $k = A+1, A+2, \ldots, B$ , the following gives a direct expression for $\boldsymbol{\mathfrak{L}}_B(.)$:*

$$\boldsymbol{\mathfrak{L}}_B(Z_i) = \left[ [1-\lambda_{A:B}]\, \boldsymbol{\mathfrak{L}}_A(Z_i) + \frac{M_i \lambda_{A:B}}{H} \sum_{j=1}^{V} \boldsymbol{\mathfrak{L}}_A(Z_j) \right] \times P(S_B|Z_i)$$

*Proof.* We will use mathematical induction to prove this claim. Proving it for $B = A+1$ follows easily from the fact that $\lambda_{A:A+1} = \lambda_A$ (substitute $B = A+1$ in Eq. (4.1), p.42). To prove it for general $B > A$ (under mathematical induction hypothesis), we assume its true for some $B = A+K$ and we will prove it here for $B = A+K+1$ for $K > 0$. To this end, we first note

that substituting $k = A + K + 1$ in Eq (3.2) we get:

$$
\begin{aligned}
& \boldsymbol{\mathfrak{L}}_{A+K+1}(Z_i) \\
= & \left[ [1-\lambda_{A+K}]\, \boldsymbol{\mathfrak{L}}_{A+K}(Z_i) + \frac{M_i \lambda_{A+K}}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A+K}(Z_j) \right] \times P(S_{A+K+1}|Z_i) \\
= & [1-\lambda_{A+K}]\, \boldsymbol{\mathfrak{L}}_{A+K}(Z_i) + \frac{M_i \lambda_{A+K}}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A+K}(Z_j) \\
= & [[1-\lambda_{A+K}]\, \boldsymbol{\mathfrak{L}}_{A+K}(Z_i) + \frac{M_i \lambda_{A+K}}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A}(Z_j)] \\
= & [1-\lambda_{A+K}] \left[ (1-\lambda_{A:A+K})\, \boldsymbol{\mathfrak{L}}_{A}(Z_i) + \frac{M_i \lambda_{A:A+K}}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A}(Z_j) \right] \\
& \qquad + \frac{M_i \lambda_{A+K}}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A}(Z_j) \\
= & (1-\lambda_{A+K})(1-\lambda_{A:A+K})\, \boldsymbol{\mathfrak{L}}_{A}(Z_i) \\
& \qquad + (\lambda_{A:A+K}(1-\lambda_{A+K}) + \lambda_{A+K}) \left[ \frac{M_i}{H} \right] \left[ \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A}(Z_j) \right] \\
= & (1-\lambda_{A:A+K+1})\, \boldsymbol{\mathfrak{L}}_{A}(Z_i) \\
& \qquad + (\lambda_{A+K} + \lambda_{A:A+K} - \lambda_{A+K}\lambda_{A:A+K}) \left[ \frac{M_i}{H} \right] \left[ \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A}(Z_j) \right] \\
= & (1-\lambda_{A:A+K+1})\, \boldsymbol{\mathfrak{L}}_{A}(Z_i) + \frac{M_i \lambda_{A:A+K+1}}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A}(Z_j) \\
= & (1-\lambda_{A:A+K+1})\, \boldsymbol{\mathfrak{L}}_{A}(Z_i) + \frac{M_i \lambda_{A:A+K+1}}{H} \sum_{j=1}^{U} \boldsymbol{\mathfrak{L}}_{A+k}(Z_j)
\end{aligned}
$$

The above follows from the induction hypothesis, Proposition (E.1.1), and the equation $\lambda_{A:A+K+1} = \lambda_{A+K} + \lambda_{A:A+K} - \lambda_{A+K}\lambda_{A:A+K}$ which follows from Eq. (4.1), p.42. This proves the hypothesis for $B = A + K + 1$.

□

The above equation proves that the forward equations in minimac4 (given by Eq. (4.5), (4.6), and (4.7) on page 43) are mathematically equivalent to those of minimac3 (Eq. (3.1), (3.2), and

(3.3) on page 27) which is in turn equivalent to that of minimac (Eq. (A.1)) on page 74). The same can be proved similarly for the backward equations. Together they would yield the same posterior probabilities. However, as mentioned before, these posterior probabilities can be used to generate allelic probabilities (imputed dosages) only at the genotyped markers where the states of space $\boldsymbol{\mathcal{S}}_3$ agree. They cannot produce allelic probabilities at ungenotyped markers where the states might not agree. This is the reason we transform the probabilities back to space $\boldsymbol{\mathcal{S}}_2$ to impute the ungenotyped markers.

**Proposition E.1.3.** *The transformation ($\boldsymbol{\mathfrak{L}}_Q \rightarrow \boldsymbol{\mathcal{L}}_Q$) obtained by the composition of transformations ($\boldsymbol{\mathfrak{L}}_Q \rightarrow L_Q$) (Eq. (3.3), p.27) and ($L_Q \rightarrow \boldsymbol{\mathcal{L}}_Q$) (Eq. (3.1), p.27) can be done directly by the following equation:*

$$\boldsymbol{\mathcal{L}}_Q(Y_j) = \boldsymbol{\mathfrak{L}}_Q^R(Z_i) \times \left[\frac{N_j}{M_i}\right] + \boldsymbol{\mathfrak{L}}_Q^{NR}(Z_i) \left[\frac{\boldsymbol{\mathcal{L}}_P(Y_j)}{\boldsymbol{\mathfrak{L}}_P(Z_i)}\right]$$

*Proof.*

$$\begin{aligned}
& \boldsymbol{\mathcal{L}}_Q\left(Y_j\right) \\
= & \sum_{l=1}^{H} I(X_l = Y_j) \times L_Q(X_l) \\
= & \sum_{l=1}^{H} I(X_l = Y_j) \times \left[\boldsymbol{\mathfrak{L}}_Q^R(Z_i) \times \left[\frac{1}{M_i}\right] + \boldsymbol{\mathfrak{L}}_Q^{NR}(Z_i) \left[\frac{L_P(X_l)}{\boldsymbol{\mathfrak{L}}_P(Z_i)}\right]\right] \\
= & \boldsymbol{\mathcal{L}}_Q^R(Z_i) \times \left[\frac{\sum_{l=1}^{H} I(X_l = Y_j)}{M_i}\right] + \boldsymbol{\mathcal{L}}_Q^{NR}(Z_i) \left[\frac{\sum_{l=1}^{H} I(X_l = Y_j) L_P(X_l)}{\boldsymbol{\mathfrak{L}}_P(Z_i)}\right] \\
= & \boldsymbol{\mathcal{L}}_Q^R(Z_i) \times \left[\frac{N_j}{M_i}\right] + \boldsymbol{\mathcal{L}}_Q^{NR}(Z_i) \left[\frac{\boldsymbol{\mathcal{L}}_P(Y_j)}{\boldsymbol{\mathfrak{L}}_P(Z_i)}\right]
\end{aligned}$$

□

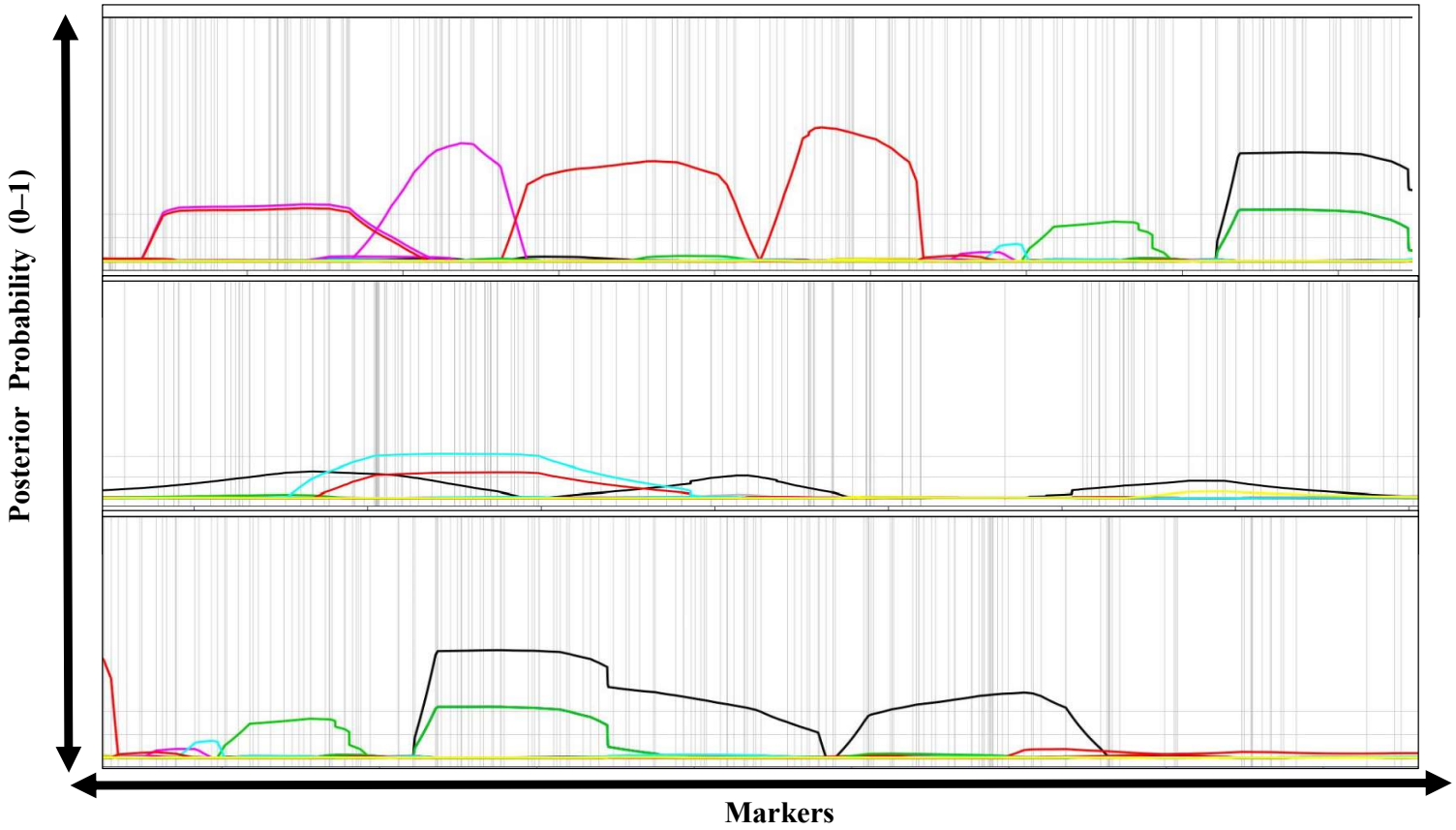


Figure E.1: Posterior probability of reference haplotypes. In this figure the the posterior probabilities of each reference haplotype is plotted for 3 different GWAS study haplotypes on chromosome 20. The X-axis represent the genomic region, the Y-axis denotes the posterior probability each reference haplotype, the colored lines each denote a reference haplotype, and the gray vertical lines denote the position of genotyped markers. The reference panel had 200 haplotypes, however, in a small genomic region only a handful of them had probability significantly larger than 0.

| | 1000G Phase 1 | | | 1000G Phase 3 | | | HRC | | | TOPMed | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MAF | M3 | M4 | # Variants | M3 | M4 | # Variants | M3 | M4 | # Variants | M3 | M4 | # Variants |
| 0.00014 | 0.1316 | 0.1295 | 169,999 | 0.1977 | 0.1951 | 213,399 | 0.4636 | 0.4647 | 204,916 | 0.5492 | 0.5499 | 193,407 |
| 0.00070 | 0.3600 | 0.3551 | 21,838 | 0.3364 | 0.3357 | 25,068 | 0.6223 | 0.6228 | 26,984 | 0.6751 | 0.6762 | 26,364 |
| 0.00141 | 0.4209 | 0.4205 | 14,656 | 0.4327 | 0.4333 | 18,236 | 0.6882 | 0.6885 | 19,412 | 0.7203 | 0.7212 | 19,131 |
| 0.00333 | 0.5622 | 0.5625 | 18,645 | 0.5732 | 0.5740 | 22,490 | 0.7789 | 0.7792 | 22,722 | 0.8011 | 0.8013 | 22,618 |
| 0.00728 | 0.6167 | 0.6167 | 16,959 | 0.6589 | 0.6593 | 17,926 | 0.8099 | 0.8099 | 17,678 | 0.8212 | 0.8204 | 17,671 |
| 0.01236 | 0.7122 | 0.7123 | 10,906 | 0.7408 | 0.7406 | 11,009 | 0.8651 | 0.8652 | 10,857 | 0.8648 | 0.8645 | 10,808 |
| 0.01745 | 0.7618 | 0.7608 | 7,779 | 0.7926 | 0.7915 | 7,789 | 0.8944 | 0.8944 | 7,685 | 0.8917 | 0.8911 | 7,668 |
| 0.02667 | 0.8261 | 0.8260 | 15,013 | 0.8451 | 0.8453 | 15,002 | 0.9126 | 0.9125 | 14,800 | 0.9110 | 0.9106 | 14,738 |
| 0.04213 | 0.8828 | 0.8824 | 9,123 | 0.8932 | 0.8927 | 9,085 | 0.9414 | 0.9410 | 8,979 | 0.9398 | 0.9390 | 8,916 |
| 0.07263 | 0.9327 | 0.9318 | 17,911 | 0.9381 | 0.9372 | 17,832 | 0.9651 | 0.9646 | 17,612 | 0.9654 | 0.9648 | 17,557 |
| 0.14518 | 0.9619 | 0.9609 | 17,576 | 0.9642 | 0.9636 | 17,503 | 0.9802 | 0.9798 | 17,252 | 0.9793 | 0.9789 | 17,204 |
| 0.24461 | 0.9718 | 0.9707 | 10,375 | 0.9736 | 0.9728 | 10,350 | 0.9858 | 0.9853 | 10,207 | 0.9850 | 0.9845 | 10,167 |
| 0.34783 | 0.9737 | 0.9726 | 6,795 | 0.9772 | 0.9765 | 6,780 | 0.9867 | 0.9863 | 6,679 | 0.9866 | 0.9862 | 6,629 |
| 0.44569 | 0.9752 | 0.9747 | 5,081 | 0.9776 | 0.9771 | 5,077 | 0.9869 | 0.9864 | 4,985 | 0.9857 | 0.9853 | 4,969 |
| 0.54873 | 0.9750 | 0.9740 | 3,760 | 0.9757 | 0.9750 | 3,742 | 0.9850 | 0.9841 | 3,709 | 0.9843 | 0.9838 | 3,666 |
| 0.64823 | 0.9739 | 0.9724 | 2,958 | 0.9769 | 0.9760 | 2,953 | 0.9878 | 0.9873 | 2,918 | 0.9879 | 0.9878 | 2,899 |
| 0.74896 | 0.9698 | 0.9695 | 1,961 | 0.9748 | 0.9747 | 1,954 | 0.9849 | 0.9845 | 1,932 | 0.9816 | 0.9814 | 1,927 |
| 0.84390 | 0.9660 | 0.9658 | 1,351 | 0.9685 | 0.9674 | 1,344 | 0.9825 | 0.9824 | 1,338 | 0.9790 | 0.9788 | 1,342 |
| 0.96378 | 0.9296 | 0.9291 | 1,776 | 0.9295 | 0.9288 | 1,822 | 0.9674 | 0.9663 | 1,661 | 0.9548 | 0.9527 | 1,867 |

Table E.1: Imputation accuracy of minimac4 compared to minimac3. This tables summarizes the imputation accuracy of minimac3 (M3) and minimac4 (M4) for imputing into 10 CG samples using 1000G Phase 1, 1000G Phase 3, HRC, and TOPMed reference panels. Minimac4 uses approximation for imputing non-genotyped sites but the accuracy falls negligibly compared to minimac3

## APPENDIX F

### Principal components of 1,000 randomly selected TOPMed samples

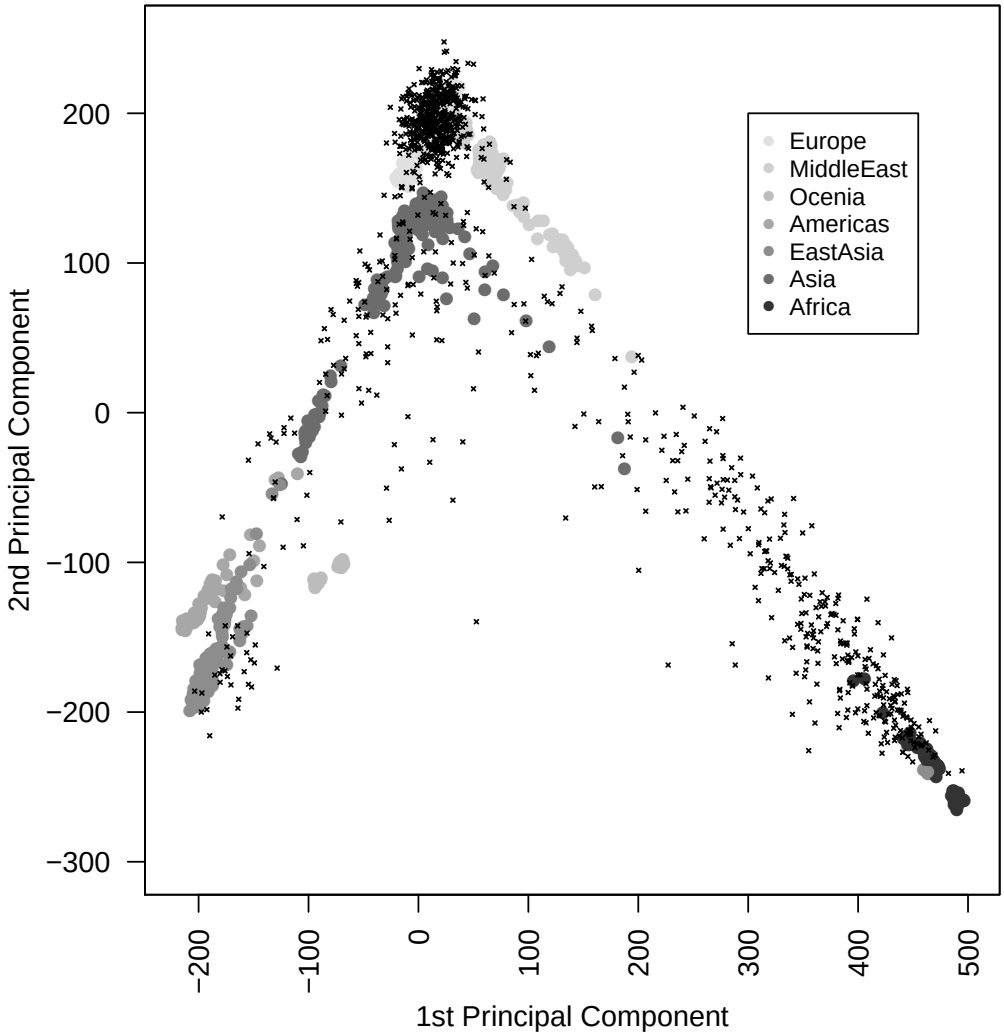


Figure F.1: Ancestry of 1,000 TOPMed samples. The figure shows the principal components of our pseudo-GWA dataset. Over ~50% of the samples are European, ~30% are African, and less than ~10% are Asian or East Asian.

# APPENDIX G

## Supplementary tables/figures for meta-imputation

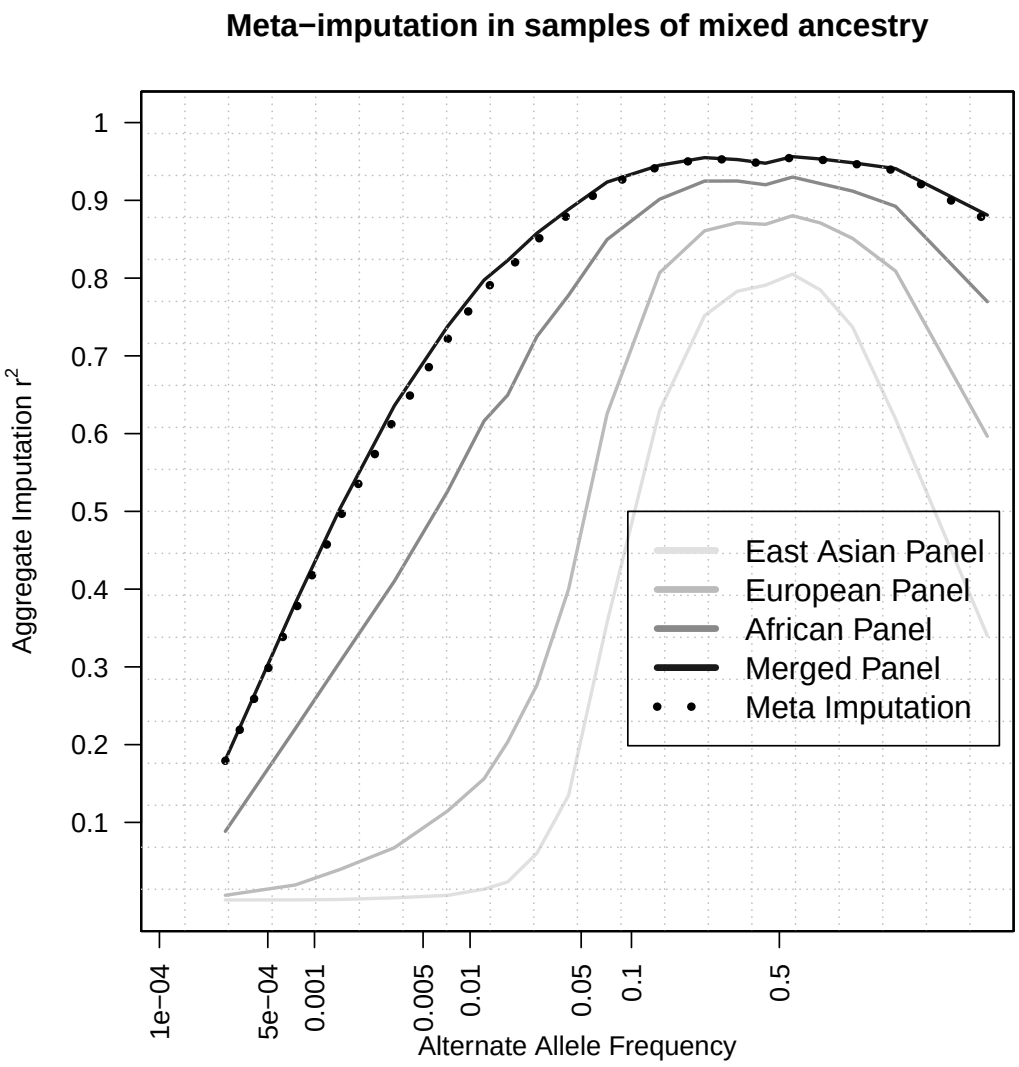


Figure G.1: Meta-imputation in samples of mixed ancestry. The figure shows the imputation accuracy of different panels and methods in samples of mixed ancestry. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$.

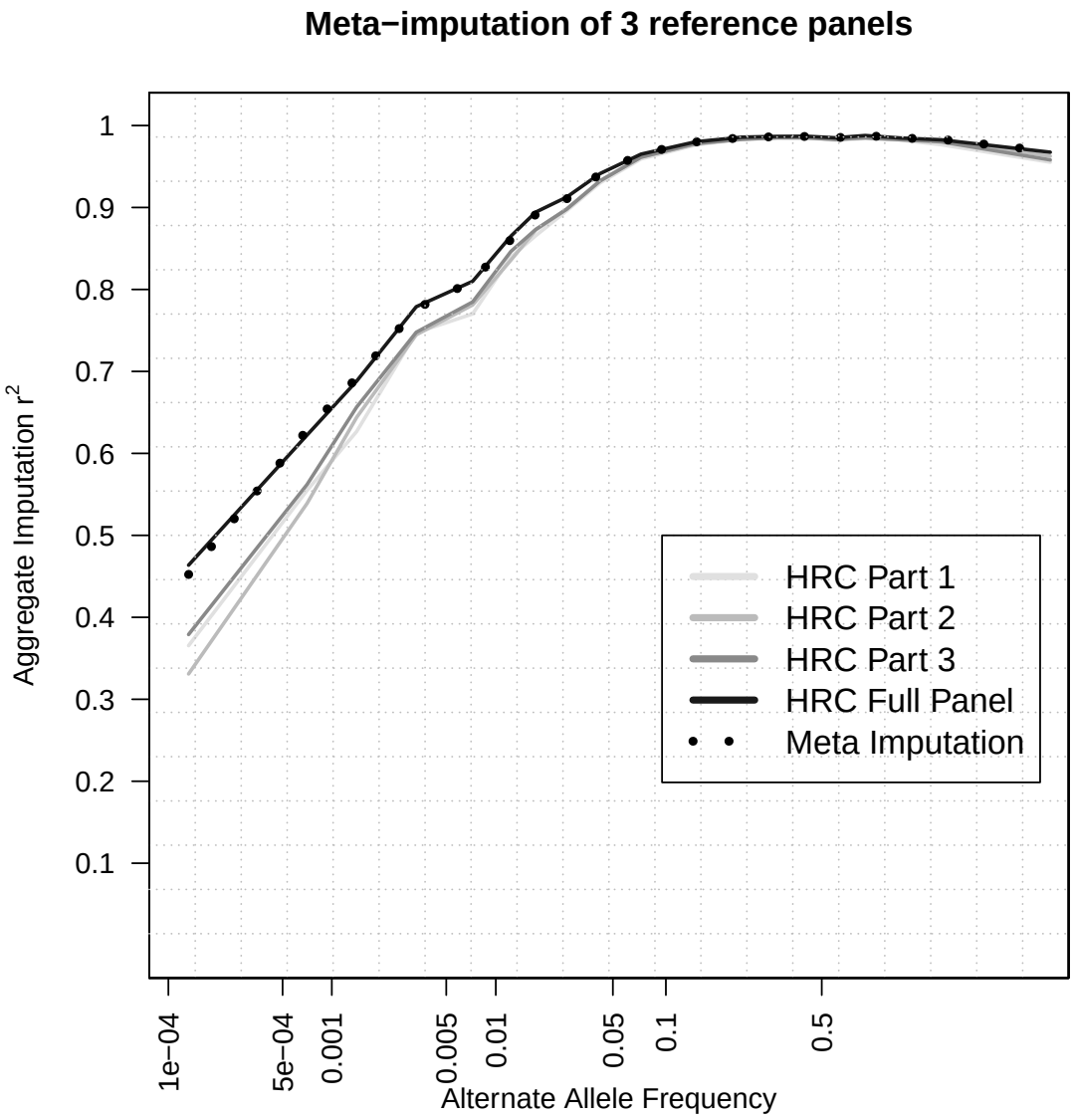


Figure G.2: Meta-imputation of 3 reference panels. The figure shows the imputation accuracy by meta-imputing the HRC reference panel in 3 non-overlapping parts. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$.

Table G.1: Imputation accuracy details for meta-imputation. This table details the aggregate $r^2$ for joint and meta-imputation for three different experiments: imputing African American samples using EUR and AFR (Section 5.3.1), imputing 1,000 samples of mixed ancestry using EUR, AFR, and EAS (Section 5.3.2), and imputing 10 European samples by splitting HRC panel into three non-overlapping parts (Section 5.3.3).

| | | African American | | Mixed Ancestry | | HRC Split Study | |
|---|---|---|---|---|---|---|---|
| Alternate AF | # Variants | Merged Panel | Meta-Imputation | Merged Panel | Meta-Imputation | Full HRC Panel | Meta-Imputation |
| 0.0003 | 117,467 | 0.058 | 0.059 | 0.180 | 0.179 | 0.464 | 0.461 |
| 0.0008 | 70,313 | 0.421 | 0.423 | 0.382 | 0.373 | 0.622 | 0.627 |
| 0.0015 | 83,780 | 0.567 | 0.564 | 0.503 | 0.492 | 0.688 | 0.703 |
| 0.0033 | 108,685 | 0.674 | 0.672 | 0.636 | 0.619 | 0.779 | 0.785 |
| 0.0072 | 73,747 | 0.768 | 0.765 | 0.738 | 0.722 | 0.810 | 0.815 |
| 0.0123 | 38,626 | 0.818 | 0.815 | 0.798 | 0.785 | 0.865 | 0.867 |
| 0.0173 | 23,542 | 0.825 | 0.823 | 0.823 | 0.811 | 0.894 | 0.895 |
| 0.0265 | 38,437 | 0.862 | 0.859 | 0.858 | 0.849 | 0.913 | 0.914 |
| 0.0418 | 19,099 | 0.874 | 0.870 | 0.888 | 0.882 | 0.941 | 0.942 |
| 0.0718 | 31,848 | 0.899 | 0.898 | 0.923 | 0.920 | 0.965 | 0.966 |
| 0.1445 | 30,817 | 0.920 | 0.920 | 0.945 | 0.944 | 0.980 | 0.981 |
| 0.2478 | 18,197 | 0.929 | 0.929 | 0.955 | 0.954 | 0.986 | 0.986 |
| 0.3488 | 14,016 | 0.930 | 0.931 | 0.952 | 0.952 | 0.987 | 0.987 |
| 0.4486 | 11,046 | 0.926 | 0.928 | 0.948 | 0.947 | 0.987 | 0.987 |
| 0.5488 | 9,294 | 0.933 | 0.934 | 0.956 | 0.955 | 0.985 | 0.986 |
| 0.6481 | 7,327 | 0.928 | 0.929 | 0.953 | 0.952 | 0.988 | 0.989 |
| 0.7481 | 5,664 | 0.930 | 0.930 | 0.948 | 0.947 | 0.985 | 0.985 |
| 0.8490 | 4,325 | 0.917 | 0.918 | 0.941 | 0.939 | 0.983 | 0.982 |

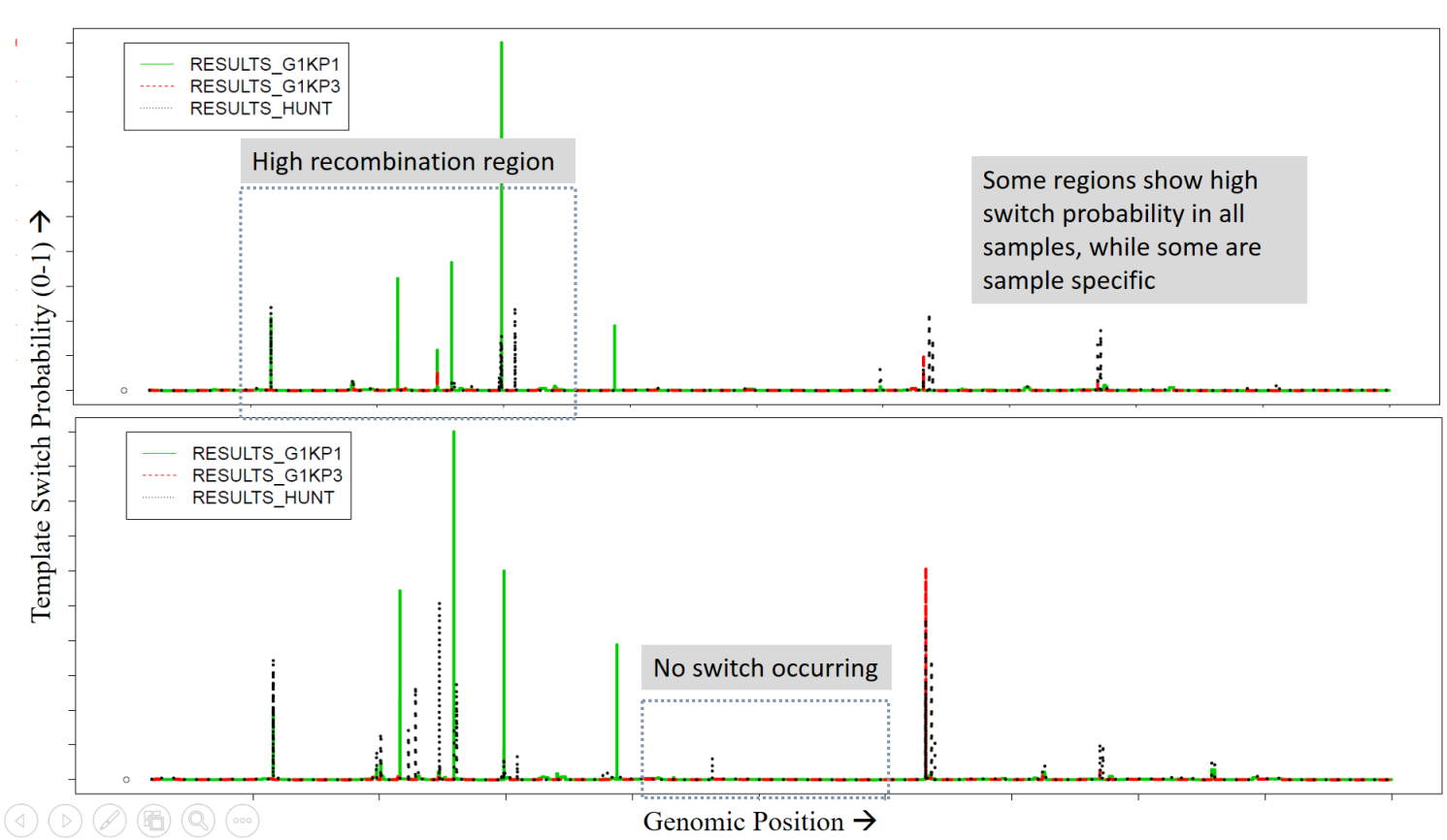


Figure G.3: Plot of template switch probability across the chromosome. The figure shows the plot between consecutive markers for two sample (one in each panel). Three reference panels 1000G Phase 1 (abbreviated as G1KP1), 1000G Phase 3 (G1KP3) and HUNT were used for imputation. Long stretches of low template switch probability denote the same haplotype from the panel was being copied over that stretch. High spikes denote regions with high chances of recombination events (template switches). The HUNT panel had the smallest average number of template switches occurring, which is as expected because genetic similarity would show the longest stretches of haplotype sharing

## APPENDIX H

### Imputation accuracy graphs for samples of non-European ancestry

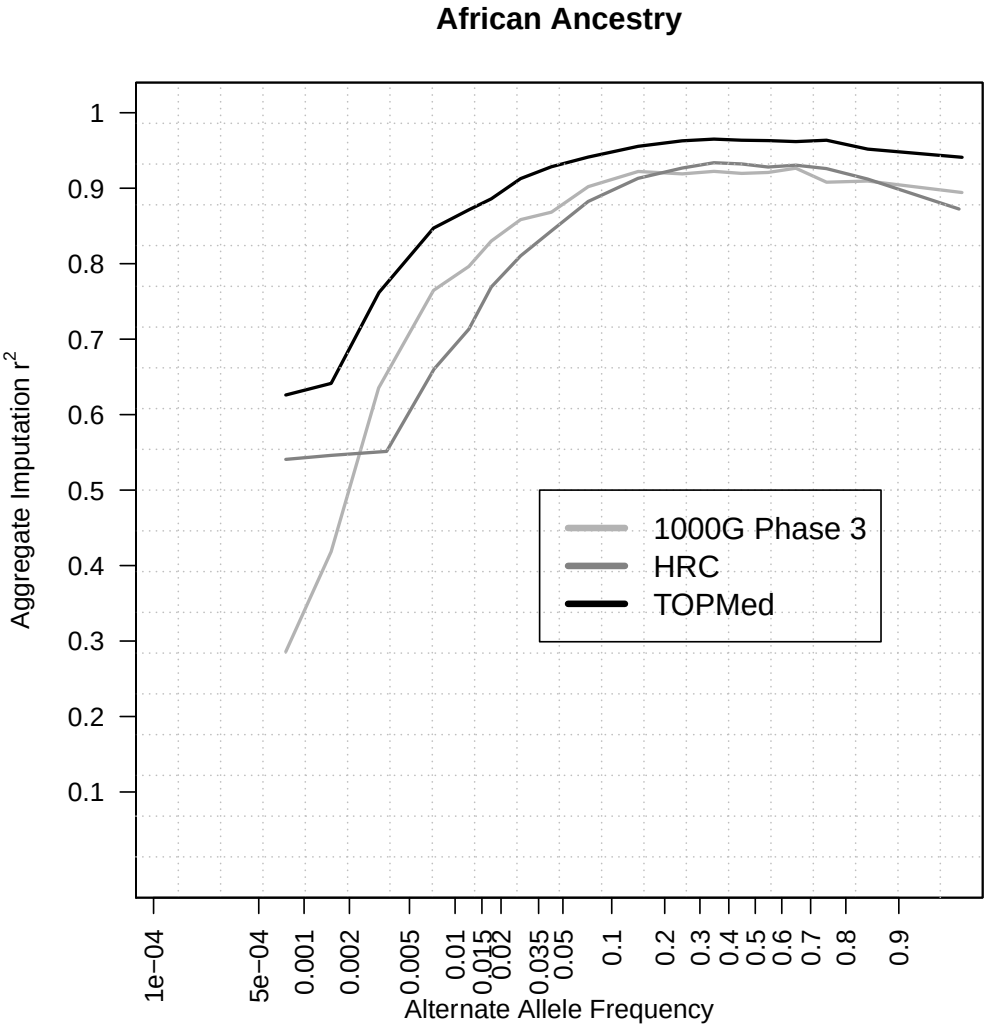


Figure H.1: Imputation accuracy in African American samples. The figure shows the imputation accuracy in samples of African American Ancestry. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$. The gray lines represent accuracy profiles for different reference panels.

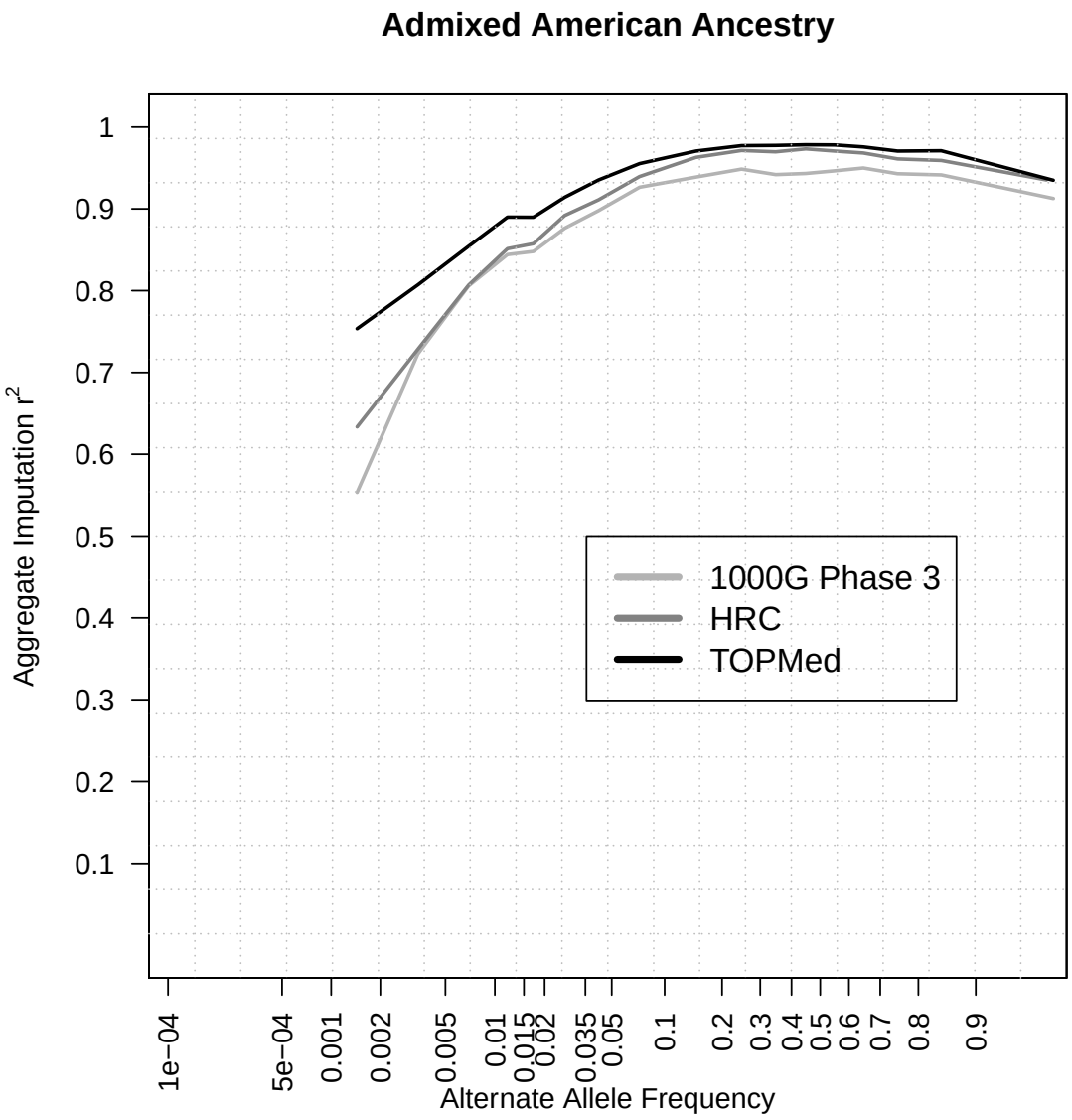


Figure H.2: Imputation accuracy in admixed American samples. The figure shows the imputation accuracy in samples of admixed American Ancestry. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$. The gray lines represent accuracy profiles for different reference panels.

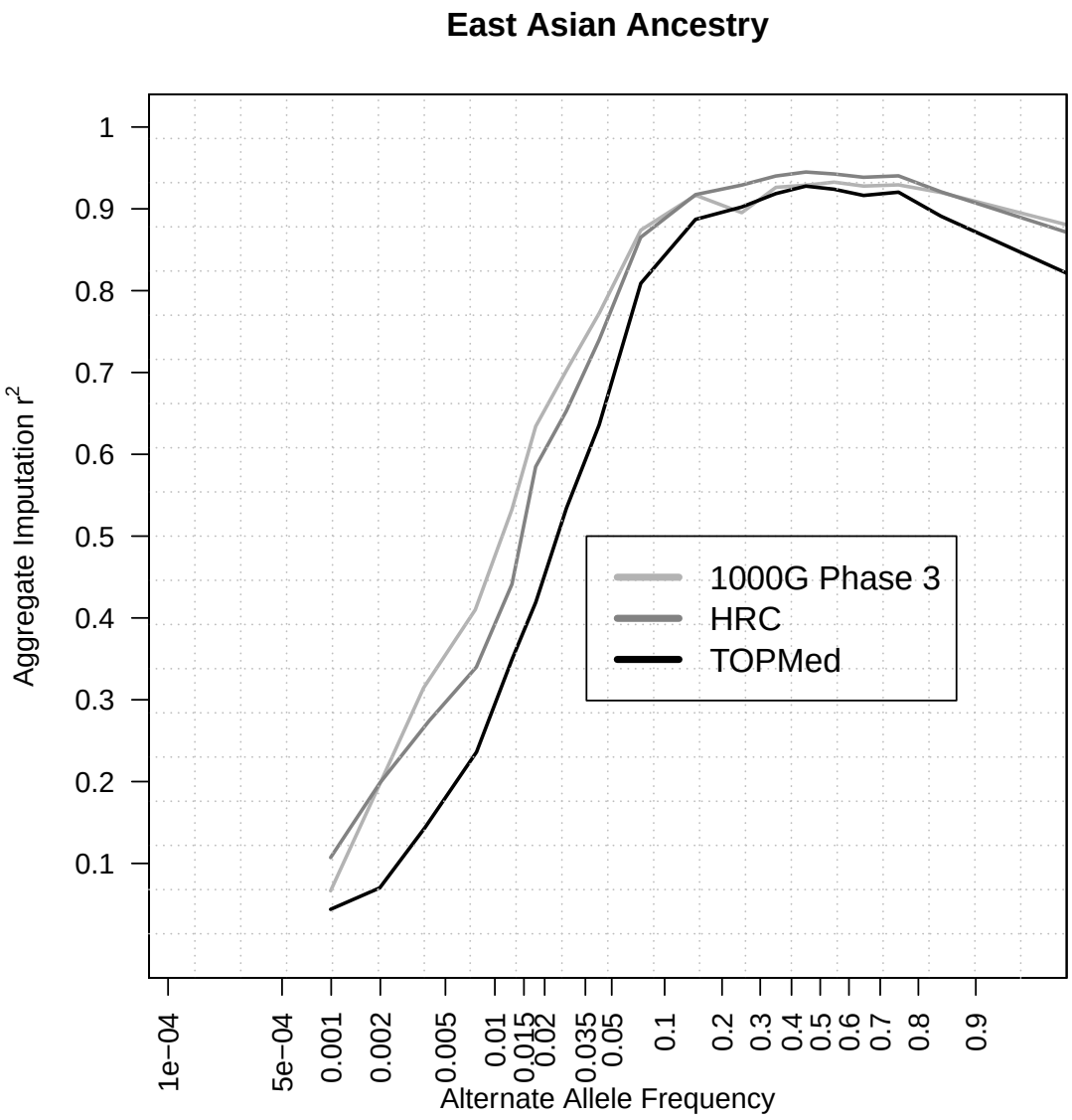


Figure H.3: Imputation accuracy in East Asian samples. The figure shows the imputation accuracy in samples of East Asian Ancestry. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$. The gray lines represent accuracy profiles for different reference panels.

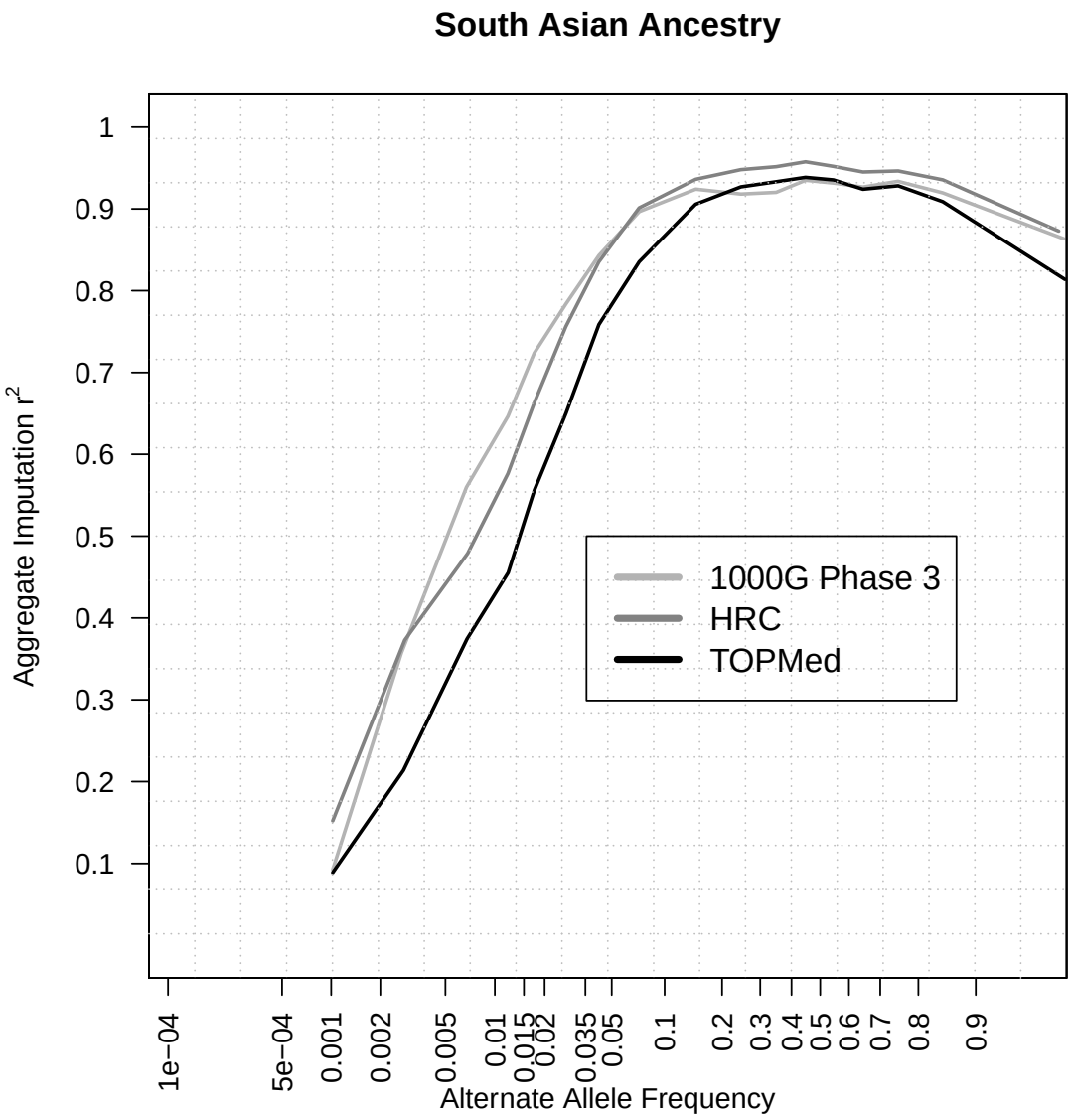


Figure H.4: Imputation accuracy in South Asian samples. The figure shows the imputation accuracy in samples of South Asian Ancestry. The X-axis shows the alternate allele frequency on a log scale. The Y-axis shows imputation accuracy measured by aggregate $r^2$. The gray lines represent accuracy profiles for different reference panels.

# APPENDIX I

## Supplementary figures for the Michigan imputation server

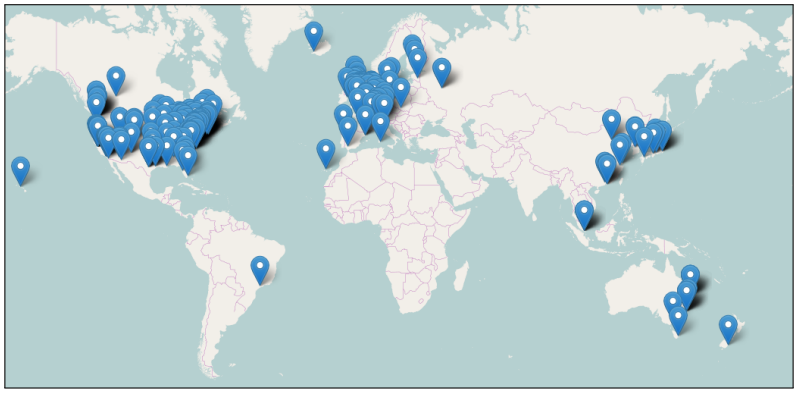

Figure I.1: Geo-location of users of Michigan imputation server

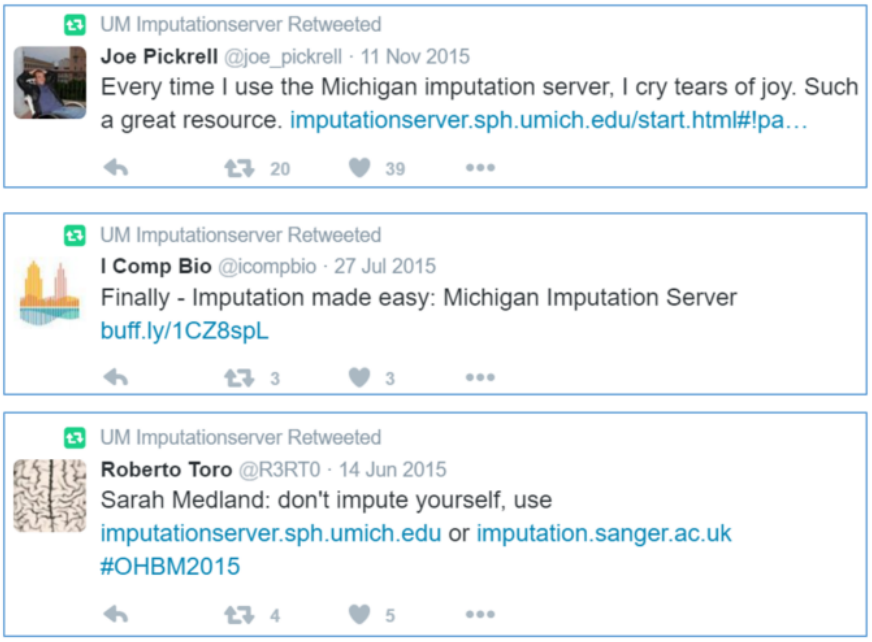


Figure I.2: Tweets by some users on the Michigan imputation server

# Bibliography

1000G, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, et al. (2010). "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319, pp. 1061–1073 (cit. on p. 7).

1000G, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65 (cit. on pp. 7, 24).

1000G et al. (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74 (cit. on pp. 7, 12, 17, 24, 81).

Alioto, T. S., I. Buchhalter, S. Derdak, B. Hutter, M. D. Eldridge, et al. (2015). "A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing". In: *Nature Communications* 6, p. 10001 (cit. on p. 2).

Amdahl, G. M. (1967). "Validity of the single processor approach to achieving large scale computing capabilities". In: *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*. ACM, pp. 483–485 (cit. on p. 50).

Anderson, C. A., F. H. Pettersson, J. C. Barrett, J. J. Zhuang, J. Ragoussis, et al. (2008). "Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms". In: *The American Journal of Human Genetics* 83.1, pp. 112–119 (cit. on p. 4).

Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, et al. (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease". In: *Nature Genetics* 40.8, pp. 955–962 (cit. on pp. 2, 81).

Bathurst, I. C., J. Travis, P. M. George, and R. W. Carrell (1984). "Structural and functional characterization of the abnormal Z alpha 1-antitrypsin isolated from human liver". In: *FEBS Letters* 177.2, pp. 179–183 (cit. on p. 23).

Baum, L. E. and T. Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains". In: *The Annals of Mathematical Statistics* 37.6, pp. 1554–1563 (cit. on p. 73).

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains”. In: *The Annals of Mathematical Statistics* 41.1, pp. 164–171 (cit. on p. 73).

Browning, B. L. and S. R. Browning (2016). “Genotype Imputation with Millions of Reference Samples”. In: *The American Journal of Human Genetics* 98.1, pp. 116–126 (cit. on pp. 7, 9–11, 13, 30, 31).

Browning, S. R. (2008). “Missing data imputation and haplotype phase inference for genome-wide association studies”. In: *Human Genetics* 124.5, pp. 439–450 (cit. on pp. 7, 9).

Chanda, P., N. Yuhki, M. Li, J. S. Bader, A. Hartz, et al. (2012). “Comprehensive evaluation of imputation performance in African Americans”. In: *Journal of Human Genetics* 57.7, pp. 411–421 (cit. on p. 64).

Chen, W., D. Stambolian, A. O. Edwards, K. E. Branham, M. Othman, et al. (2010). “Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration”. In: *Proceedings of the National Academy of Sciences* 107.16, pp. 7401–7406 (cit. on p. 81).

Colonna, V., G. Pistis, L. Bomba, S. Mona, G. Matullo, et al. (2012). “Small effective population size and genetic homogeneity in the Val Borbera isolate”. In: *European Journal of Human Genetics* 21.1, pp. 89–94 (cit. on p. 81).

CONVERGE (2015). “Sparse whole-genome sequencing identifies two loci for major depressive disorder”. In: *Nature* 523.7562, pp. 588–591 (cit. on p. 23).

Coon, K. D., A. J. Myers, D. W. Craig, J. A. Webster, J. V. Pearson, et al. (2007). “A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer’s disease”. In: *The Journal of Clinical Psychiatry* 68.4, pp. 613–618 (cit. on p. 2).

Cooper, J. D., D. J. Smyth, A. M. Smiles, V. Plagnol, N. M. Walker, et al. (2008). “Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci”. In: *Nature Genetics* 40.12, pp. 1399–1401 (cit. on p. 4).

Das, S., L. Forer, S. Schonherr, C. Sidore, A. E. Locke, et al. (2016). “Next-generation genotype imputation service and methods”. In: *Nature Genetics* 48.10, pp. 1284–1287 (cit. on pp. 5, 7, 10, 12, 13, 20, 38, 69).

Deelen, P., A. Menelaou, E. M. van Leeuwen, A. Kanterakis, F. van Dijk, et al. (2014). “Improved imputation quality of low-frequency and rare variants in European samples using the ‘Genome of The Netherlands’”. In: *European Journal of Human Genetics: EJHG* 22.11, pp. 1321–1326 (cit. on pp. 13, 59).

Delaneau, O., J.-F. Zagury, and J. Marchini (2012). "Improved whole-chromosome phasing for disease and population genetic studies". In: *Nature Methods* 10.1, pp. 5–6 (cit. on p. 19).

Duan, Q., E. Y. Liu, P. L. Auer, G. Zhang, E. M. Lange, et al. (2013). "Imputation of coding variants in African Americans: better performance using data from the exome sequencing project". In: *Bioinformatics (Oxford, England)* 29.21, pp. 2744–2749 (cit. on p. 14).

Dudbridge, F. (2008). "Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data". In: *Human Heredity* 66.2, pp. 87–98 (cit. on p. 9).

Durbin, R. (2014). "Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)". In: *Bioinformatics* 30.9, pp. 1266–1272 (cit. on pp. 9, 20, 38).

Esko, T., M. Mezzavilla, M. Nelis, C. Borel, T. Debniak, et al. (2012). "Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity". In: *European Journal of Human Genetics* 21.6, pp. 659–665 (cit. on p. 81).

Ferrarotti, I., G. A. Thun, M. Zorzetto, S. Ottaviani, M. Imboden, et al. (2012). "Serum levels and genotype distribution of alpha1-antitrypsin in the general population". In: *Thorax* 67.8, pp. 669–674 (cit. on p. 23).

Ferrucci, L., S. Bandinelli, E. Benvenuti, A. Di Iorio, C. Macchi, et al. (2000). "Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study". In: *Journal of the American Geriatrics Society* 48.12, pp. 1618–1625 (cit. on pp. 20, 81).

Francalacci, P., L. Morelli, A. Angius, R. Berutti, F. Reinier, et al. (2013). "Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny". In: *Science* 341.6145, pp. 565–569 (cit. on pp. 17, 24, 57).

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164 (cit. on p. 17).

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds (2014). "minimac2: faster genotype imputation". In: *Bioinformatics* 31.5, pp. 782–784 (cit. on pp. 7, 11, 24, 25, 30, 69).

Fuchsberger, C., J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, et al. (2016). "The genetic architecture of type 2 diabetes". In: *Nature* 536.7614, pp. 41–47 (cit. on p. 81).

Gadegbeku, C. A., D. S. Gipson, L. B. Holzman, A. O. Ojo, P. X. K. Song, et al. (2013). "Design of the Nephrotic Syndrome Study Network (NEPTUNE) to evaluate primary glomerular nephropathy by a multidisciplinary approach". In: *Kidney International* 83.4, pp. 749–756 (cit. on p. 81).

Ghahramani, Z. (2001). “An introduction to hidden Markov models and Bayesian networks”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 15.01, pp. 9–42 (cit. on p. 73).

Gibson, G. (2012). “Rare and common variants: twenty arguments”. In: *Nature Reviews Genetics* 13.2, pp. 135–145 (cit. on p. 2).

Global Lipids, C. J. Willer, E. M. Schmidt, S. Sengupta, G. M. Peloso, et al. (2013). “Discovery and refinement of loci associated with lipid levels”. In: *Nature Genetics* 45.11, pp. 1274–1283 (cit. on p. 1).

GoNL et al. (2014). “Whole-genome sequence variation, population structure and demographic history of the Dutch population”. In: *Nature Genetics* 46.8, pp. 818–825 (cit. on pp. 17, 24, 57, 81).

Goodwin, S., J. D. McPherson, and W. R. McCombie (2016). “Coming of age: ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17.6, pp. 333–351 (cit. on p. 2).

Gudbjartsson, D. F., H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, et al. (2015). “Large-scale whole-genome sequencing of the Icelandic population”. In: *Nature Genetics* 47.5, pp. 435–444 (cit. on pp. 24, 57).

Gurdasani, D., T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, et al. (2014). “The African Genome Variation Project shapes medical genetics in Africa”. In: *Nature* 517.7534, pp. 327–332 (cit. on p. 23).

Handsaker, R. E., V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, et al. (2015). “Large multiallelic copy number variations in humans”. In: *Nature Genetics* 47.3, pp. 296–303 (cit. on p. 4).

Hao, K., E. Chudin, J. McElwee, and E. E. Schadt (2009). “Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies”. In: *BMC Genetics* 10.1, p. 27 (cit. on p. 4).

HapMap (2003). “The International HapMap Project”. In: *Nature* 426.6968, pp. 789–796 (cit. on p. 7).

— (2005). “A haplotype map of the human genome”. In: *Nature* 437.7063, pp. 1299–1320 (cit. on p. 7).

HapMap, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, et al. (2007). “A second generation human haplotype map of over 3.1 million SNPs”. In: *Nature* 449.7164, pp. 851–861 (cit. on p. 7).

HapMap, D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler, et al. (2010). "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467.7311, pp. 52–58 (cit. on pp. 7, 12).

Hays, J., J. R. Hunt, F. A. Hubbell, G. L. Anderson, M. Limacher, et al. (2003). "The Women's Health Initiative recruitment methods and results". In: *Annals of Epidemiology* 13.9, pp. 18–77 (cit. on p. 81).

Hoffmann, T. J., M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino, et al. (2011). "Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array". In: *Genomics* 98.2, pp. 79–89 (cit. on p. 2).

Hoffmann, T. J., L. C. Sakoda, L. Shen, E. Jorgenson, L. A. Habel, et al. (2015). "Imputation of the rare HOXB13 G84E mutation and cancer risk in a large population-based cohort". In: *PLOS Genetics* 11.1. Ed. by P. M. Visscher, p. 30 (cit. on pp. 4, 14).

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing". In: *Nature Genetics* 44.8, pp. 955–959 (cit. on pp. 7, 10, 11, 24, 25, 27, 29, 30, 69).

Howie, B. N., P. Donnelly, and J. Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies". In: *PLoS Genetics* 5.6. Ed. by N. J. Schork, p. 529 (cit. on pp. 7, 14).

Huang, J., B. Howie, S. McCarthy, Y. Memari, K. Walter, et al. (2015). "Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel". In: *Nature communications* 6, pp. 457–470 (cit. on pp. 14, 17, 19, 58).

Huang, L., Y. Li, A. B. Singleton, J. A. Hardy, G. Abecasis, et al. (2009). "Genotype-imputation accuracy across worldwide human populations". In: *The American Journal of Human Genetics* 84.2, pp. 235–250 (cit. on p. 58).

HumanGenome (2004). "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011, pp. 931–945 (cit. on p. 7).

Jewett, E. M., M. Zawistowski, N. A. Rosenberg, and S. Zöllner (2012). "A coalescent model for genotype imputation". In: *Genetics* 191.4, pp. 1239–1255 (cit. on p. 59).

Jia, X., B. Han, S. Onengut-Gumuscu, W.-M. Chen, P. J. Concannon, et al. (2013). "Imputing amino acid polymorphisms in human leukocyte antigens". In: *PLoS ONE* 8.6. Ed. by J. Tang, p. 83 (cit. on p. 4).

Klein, R. J., C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, et al. (2005). "Complement factor H polymorphism in age-related macular degeneration". In: *Science* 308.5720, pp. 385–389 (cit. on pp. 1, 7).

Krokstad, S., A. Langhammer, K. Hveem, T. L. Holmen, K. Midthjell, et al. (2012). "Cohort Profile: the HUNT Study, Norway". In: *International Journal of Epidemiology* 42.4, pp. 968–977 (cit. on pp. 58, 81).

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, et al. (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height". In: *Nature* 467.7317, pp. 832–838 (cit. on p. 1).

Leslie, S., P. Donnelly, and G. McVean (2008). "A statistical method for predicting classical HLA alleles from SNP data". In: *The American Journal of Human Genetics* 82.1, pp. 48–56 (cit. on p. 4).

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al. (2009a). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics (Oxford, England)* 25.16, pp. 2078–2079 (cit. on p. 18).

Li, N. and M. Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". In: *Genetics* 165.4, pp. 2213–2233 (cit. on pp. 8, 10).

Li, Y., C. Willer, S. Sanna, and G. Abecasis (2009b). "Genotype imputation". In: *Annual Review of Genomics and Human Genetics* 10.1, pp. 387–406 (cit. on pp. 3, 12, 24).

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis (2010). "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes". In: *Genetic Epidemiology* 34.8, pp. 816–834 (cit. on pp. 7, 25, 27–29, 69, 73, 74).

Lin, D. Y., Y. Hu, and B. E. Huang (2008). "Simple and efficient analysis of disease association with missing genotype data". In: *American Journal of Human Genetics* 82.2, pp. 444–452 (cit. on p. 9).

Locke, A. E., B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, et al. (2015). "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518.7538, pp. 197–206 (cit. on pp. 1, 2).

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, et al. (2012). "A systematic survey of loss-of-function variants in human protein-coding genes". In: *Science* 335.6070, pp. 823–828 (cit. on p. 24).

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, et al. (2009). "Finding the missing heritability of complex diseases". In: *Nature* 461.7265, pp. 747–753 (cit. on p. 2).

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes". In: *Nature Genetics* 39.7, pp. 906–913 (cit. on pp. 3, 7, 29).

Marchini, J. and B. Howie (2010). “Genotype imputation for genome-wide association studies”. In: *Nature Reviews Genetics* 11.7, pp. 499–511 (cit. on pp. 3, 24).

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, et al. (2008). “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. In: *Nature reviews genetics* 9.5, pp. 356–369 (cit. on p. 1).

McCarthy, S., S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, et al. (2016). “A reference panel of 64,976 haplotypes for genotype imputation”. In: *Nature Genetics* 48.10, pp. 1279–1283 (cit. on pp. 5, 7, 12).

McQuillan, R., A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, et al. (2008). “Runs of homozygosity in European populations”. In: *The American Journal of Human Genetics* 83.3, pp. 359–372 (cit. on p. 81).

MIGCI, N. O. Stitziel, H.-H. Won, A. C. Morrison, G. M. Peloso, et al. (2014). “Inactivating mutations in NPC1L1 and protection from coronary heart disease”. In: *New England Journal of Medicine* 371.22, pp. 2072–2082 (cit. on p. 24).

Nair, R. P., K. C. Duffin, C. Helms, J. Ding, P. E. Stuart, et al. (2009). “Genomewide Scan Reveals Association of Psoriasis with IL-23 and NF-$\kappa$B Pathways”. In: *Nature Genetics* 41.2, pp. 199–204 (cit. on p. 2).

Nelder, J. A. and R. Mead (1965). “A simplex method for function minimization”. In: *The computer journal* 7.4, pp. 308–313 (cit. on p. 64).

NHLBI (2015). “Trans-Omics for Precision Medicine (TOPMed) Program” (cit. on p. 12).

Nicolae, D. L. (2006). “Testing untyped alleles (TUNA)-applications to genome-wide association studies”. In: *Genetic Epidemiology* 30.8, pp. 718–727 (cit. on p. 9).

Okada, Y., D. Wu, G. Trynka, T. Raj, C. Terao, et al. (2013). “Genetics of rheumatoid arthritis contributes to biology and drug discovery”. In: *Nature* 506.7488, pp. 376–381 (cit. on p. 2).

Orho-Melander, M., O. Melander, C. Guiducci, P. Perez-Martinez, D. Corella, et al. (2008). “Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations”. In: *Diabetes* 57.11, pp. 3112–3121 (cit. on p. 3).

Panoutsopoulou, K., K. Hatzikotoulas, D. K. Xifara, V. Colonna, A.-E. Farmaki, et al. (2014). “Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants”. In: *Nature Communications* 5, p. 5345 (cit. on p. 81).

Pato, M. T., J. L. Sobell, H. Medeiros, C. Abbott, B. M. Sklar, et al. (2013). "The genomic psychiatry cohort: partners in discovery". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 162B.4, pp. 306–312 (cit. on p. 81).

Pistis, G., E. Porcu, S. I. Vrieze, C. Sidore, M. Steri, et al. (2014). "Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs". In: *European Journal of Human Genetics* 23.7, pp. 975–983 (cit. on pp. 13, 59).

Pritchard, J. K. and M. Przeworski (2001). "Linkage disequilibrium in humans: models and data". In: *The American Journal of Human Genetics* 69.1, pp. 1–14 (cit. on p. 32).

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3, pp. 559–575 (cit. on p. 9).

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, et al. (2002). "Genetic structure of human populations". In: *Science* 298.5602, pp. 2381–2385 (cit. on p. 23).

Saunders, E. J., T. Dadaev, D. A. Leongamornlert, S. Jugurnauth-Little, M. Tymrakiewicz, et al. (2014). "Fine-mapping the HOXB region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer". In: *PLoS Genetics* 10.2. Ed. by G. Gibson, p. 29 (cit. on p. 14).

Scheet, P. and M. Stephens (2006). "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase". In: *The American Journal of Human Genetics* 78.4, pp. 629–644 (cit. on pp. 7, 8).

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, et al. (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants". In: *Science* 316.5829, pp. 1341–1345 (cit. on pp. 3, 7).

Sidore, C., F. Busonero, A. Maschio, E. Porcu, S. Naitza, et al. (2015). "Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers". In: *Nature Genetics* 47.11, pp. 1272–1281 (cit. on p. 81).

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini (2009). "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip". In: *PLoS Genetics* 5.5. Ed. by J. D. Storey, p. 77 (cit. on p. 4).

Sulem, P., H. Helgason, A. Oddson, H. Stefansson, S. A. Gudjonsson, et al. (2015). "Identification of a large set of rare complete human knockouts". In: *Nature Genetics* 47.5, pp. 448–452 (cit. on p. 24).

Tsoi, L. C., S. L. Spain, E. Ellinghaus, P. E. Stuart, F. Capon, et al. (2015). "Enhanced meta-analysis and replication studies identify five new psoriasis susceptibility loci". In: *Nature Communications* 6, p. 7001 (cit. on p. 4).

UK10K, K. Walter, J. L. Min, J. Huang, L. Crooks, et al. (2015). "The UK10K project identifies rare variants in health and disease". In: *Nature* 526.7571, pp. 82–90 (cit. on pp. 12, 81).

Vartiainen, E., T. Laatikainen, M. Peltonen, A. Juolevi, S. Männistö, et al. (2009). "Thirty-five-year trends in cardiovascular risk factors in Finland". In: *International Journal of Epidemiology* 39.2, pp. 504–518 (cit. on p. 81).

Vrieze, S. I., S. M. Malone, U. Vaidyanathan, A. Kwong, H. M. Kang, et al. (2014). "In search of rare variants: preliminary results from whole genome sequencing of 1,325 individuals with psychophysiological endophenotypes". In: *Psychophysiology* 51.12, pp. 1309–1320 (cit. on p. 81).

Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, et al. (2008). "Newly identified loci that influence lipid concentrations and risk of coronary artery disease". In: *Nature Genetics* 40.2, pp. 161–169 (cit. on p. 4).

Witte, J. S., P. M. Visscher, and N. R. Wray (2014). "The contribution of genetic variants to disease depends on the ruler". In: *Nature Reviews Genetics* 15.11, pp. 765–776 (cit. on p. 2).

WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145, pp. 661–678 (cit. on p. 1).

Yu, Z. and D. J. Schaid (2007). "Methods to impute missing genotypes for population data". In: *Human Genetics* 122.5, pp. 495–504 (cit. on p. 6).

Zhang, J., K. Jiang, L. Lv, H. Wang, Z. Shen, et al. (2015). "Use of Genome-Wide Association Studies for Cancer Research and Drug Repositioning". In: *PLOS ONE* 10.3. Ed. by G. Novelli, p. 77 (cit. on p. 2).